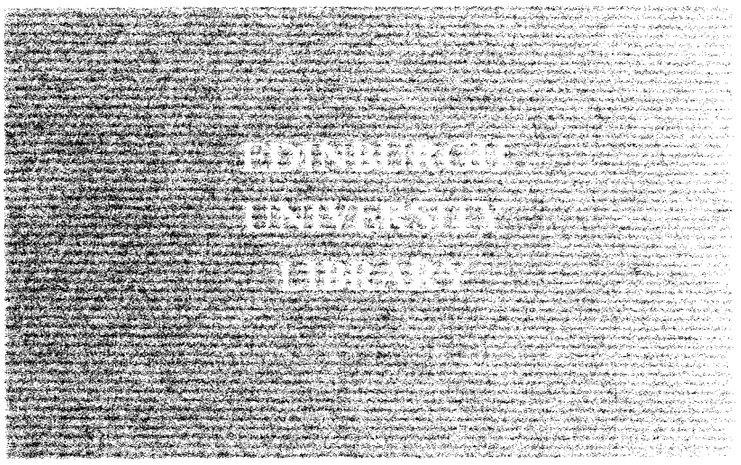

ON RULES AND THE METAPHYSICS OF MEANING

Peter Campbell

PhD

The University of Edinburgh

1999

I hereby declare that I composed this thesis myself, that it is my own work, and that it has not been submitted for any other degree or qualification.

Peter Campbell

Abstract

In this work I develop an argument which shows that rule-following is impossible, and investigate its impact on the philosophy of language. By way of orientation, I start with a critical evaluation of existing 'rule-following considerations', arguments derived from Wittgenstein which purportedly put rule-following under pressure. Having shown that its predecessors are unsound, and with the explicit aim of avoiding their flaws, I then formulate the new 'indexical' argument.

The conclusion that rule-following is impossible is difficult to accept because we think that the ability to follow rules is constitutive of language-mastery. If this is correct, then to show that rule-following is impossible is to show that language is impossible. Such 'meaning nihilism' is not a tenable position, and some way of avoiding this conclusion has to be found. Various proposals in the literature have the potential to do this: principally (a) the irrealism suggested by Kripke; and (b) subjective on-gong determination advocated by Wright. I argue that neither strategy is successful.

The correct response to the indexical argument is to accept that rule-following is not constitutive of language-mastery. In this case, clearly, the impossibility of rule-following does not entail the impossibility meaning, and the conclusion that rule-following is impossible becomes unproblematic. Nevertheless, it is difficult to see how language could survive without rules. The remainder of this work shows that rule-elimination does permit a respectable notion of linguistic content. The result is distinctively Wittgensteinian: a communitarian, 'use'-based account of language.

Contents

Introduction

This is an investigation into the nature of rule-following. It is motivated by the readily acknowledged connection between rules and meaning, namely that rules bestow words with content. To have meaning, a word must have an extension; to use the word meaningfully, I must know what that extension is. To take an example, the meaning of the English word 'red' determines that it applies to all, and only, red objects, and if I am to know the meaning of that word, I have to know that the word applies to all and only red objects. Since the word 'red' applies to countless objects - objects which I have never come across, nor even envisaged - it cannot be that I know *explicitly* which objects 'red' applies to. Rather, I must have a *principle* for using the word - that is, *a rule* governing its application to the world. It is the fact that we follow a certain rule when using a word that gives the word its meaning. As a result, the illumination of rule-following ought to allow for the illumination of meaning.

In focusing on the rule which governs the application of the word 'red' (which we may call a 'truth-rule', for it is the rule governing what the word is true of), we simplify things enormously. Complete mastery of a complex language such as English would normally be considered to involve mastery of several layers of rules. In addition to the application rules mentioned above, there are also, for example, grammatical and syntactical rules governing the way that words may be combined to form sentences and simple sentences combined to form complex ones. In addition, we follow various discourse rules which govern the way different constructions can have different senses - literal meaning as opposed to metaphorical, say. Yet it is the truth-rule that bestows a word with content, and it is the examination of these rules which promises to expose the nature of language at the most fundamental level.[1]

---

[1] Rules seem to permeate human existence at every level. Not only do we think of ourselves as following rules when we speak, but also when acting morally, lawfully, or politely. In addition, there are many wholly practical endeavours in which the avoidance of disaster appears to depend upon our satisfaction of various rules. (For example, when driving, stay on the stipulated side of the road; when constructing a building, use your materials to within their load-bearing capacity, and so on.) The ramifications of any examination of rule-following, therefore, extend far beyond the philosophy of language, having bearing on ethics, the philosophy of law, and indeed on the framework used to explain human behaviour in general. Whilst I acknowledge the wider significance that rule-following has, it is not possible to consider such matters within this work.

Wittgenstein's Attack

Whilst the content-bestowing power of rules is sufficient to make the formulation of a theory of rule-following worthwhile, the endeavour is made wholly compelling by the fertility of Wittgenstein's own investigations in this area. It is now broadly recognised that Wittgenstein's "rule-following considerations" (principally *Philosophical Investigations* §§ 139-242, and *Remarks on the Foundations of Mathematics* Part VI) lie at the heart of his later philosophy. From the examination of rules, he formulates an original philosophy of language, which in turn supports theses in the philosophies of logic, mathematics, mind, and of the nature of philosophy itself. In this way, the alteration to our views about language arising from the investigation of rules gives unprecedented potential for shedding new light on old problems.

The foundation of Wittgenstein's argument is the contention that our commonplace (philosophical) thoughts about rules and rule-following are fundamentally wrong. As a result of the negative argument we are forced to *revise* of our notion of rule-following, and, given the noted connection between rules and meaning, a corresponding alteration to our views about language ensues. Without giving any further detail, the overall structure of Wittgenstein's argument can thus be given in three parts:

Negative Argument: A demonstration that our ordinary conception of rule-following is somehow flawed.

Positive Thesis: A revision of the notion of rule-following to accommodate the conclusion of the negative argument.

Application: The incorporation of the revised theory of rule-following within a (revised) theory of meaning.

Should we try to pin down the detail of Wittgenstein's argument in more detail, we would face controversy at every stage. There is no consensus as to the identity of the 'ordinary conception' of rule-following that Wittgenstein intended to overturn; the nature of the negative argument; the positive steps he suggested should be taken to remedy the situation; the impact of these concerns on the rest of his philosophy.

It is not my aim to settle any exegetical disputes here, for it is the overall strategy which is of interest. Certainly, without needing to settle on the exact interpretation of his work, it is clear that Wittgenstein's own investigation into rule-following was original and penetrating, and

of such significance to him that it became the foundation for much of his (later) philosophy. At the outset it would be rash to suggest that Wittgenstein's thoughts on the subject can be bettered, or that his work leaves room for any substantial development of the topic. In no way do I wish to play down the magnitude of Wittgenstein's contribution, or the degree in which every discussion of rule-following, including this one, is indebted to him. However, the automatic limitation of our discussion to *Wittgenstein's* rule-following considerations is a needless restriction. Philosophy is better served, I think, if, rather than asking what Wittgenstein intended to show, we instead take this overall structure and see how far it can take us. With this aim, the questions before us are: What, if anything, is the problem with rule-following? What is the appropriate response? What impact does this have on the philosophy of language?

The Paradox of Rule-Following

My formulation of the tripartite structure given above is inspired by Saul Kripke's influential *Wittgenstein on Rules and Private Language*,[2] where the separation of negative argument from corresponding positive thesis is given particular prominence. (Although Wittgenstein did not lay out the tripartite structure given above explicitly, it is beyond question that Wittgenstein's intentions in discussing rules were polemical - not merely to elaborate on our ordinary characterisation of rules, but to *alter* our views - in which case his first step must have been to show that there is something wrong with the (then) received position. In making the division so prominent, we do not thereby misrepresent the substance of the argument, but merely make the somewhat obscure structure more visible.) Given that it is possible to run these first two steps of the argument together, to present a single unified case in support of the posited revisionary thesis, it might be thought that this division is somewhat superficial. However, in Kripke's hands the division is far from idle, for it is used to good motivational effect. (My reason for retaining this aspect of Kripke's presentation is precisely to exploit that effect here.)

The motivation in question comes from the generation of a 'paradox of rule-following'. The paradox arises when we note that the reason for rejecting the ordinary conception of rule-following is that rule-following so characterised cannot occur. In short, the argument must show that rule-following, as ordinarily characterised, is impossible. On the basis of the above noted constitutive connection between rules and meaning, if rule-following is impossible,

---

[2] Kripke (1982).

then so too is meaning. If this is right, language is illusory. Such a conclusion is certainly difficult (if not impossible) to accept,[3] for if meaning is impossible, we could not *state* the conclusion of the argument; nor, for that matter, could we *formulate* the argument in the first place. In short, the 'sceptical' argument undermines itself, which is why Kripke calls it the 'paradox' of the rule-following considerations.

(For my part I cannot accept that this 'paradox' is a genuine antimony. Clearly meaning nihilism is a position which cannot be established by argument - indeed by its own lights it cannot even be stated - but that alone does not ensure that meaning is, after all, possible. This is because a stable response to the self-defeating nature of any such argument would be to accept that it is not an argument at all *precisely because* language is not possible. This, though, is a minor point, for a paradox can also be a demonstration that what manifestly is the case cannot actually occur. The existence of a paradox of rule-following in this more informal sense is quite sufficient to galvanise a search for a disarming response to a negative argument. Meaningful communication is possible; we do it all the time. The challenge is to see how this can be so.)

To avoid the paradox, Kripke suggests that a more sophisticated response has to be found, one which avoids the self-defeating result. To this end, Kripke portrays his negative argument - what he calls the "sceptical argument" - as a step in a larger argument which ends with a retreat from realism about meanings. I shall examine Kripke's alternative conclusion in Part 2, but for now our interest lies with the effect of the paradox, which is to *force* us to reconsider rule-following at a fundamental level, to formulate some new position not previously on the map. It is precisely because any negative argument (not just Kripke's) will give rise to the same paradox, and so will similarly compel us to expand our horizons, that the search for any such argument gains its true significance.

Strategy

When I started this work I was not convinced that anyone had found a substantial problem with rule-following, which is to say that I was not convinced that the rule-following considerations had any substantial message to convey. What I was certain about was that the rule-following considerations reflect on the notion of meaning at such a fundamental level, that there is no more important question in the philosophy of language than whether any of

---

[3] Though this view - that meaningful communication is impossible - was apparently held by Cratylus.

the positions claimed as the result of this type of enquiry could in fact be established. Whether actually ascribed to Wittgenstein, or simply inspired by Wittgenstein, there is a wide range of theories offered as "the message of the rule-following considerations". Those already on the table include: that meaning is communal; that meaning is public; that meanings are not real; that meanings are non-naturalistic; that they are subjective; dispositional; unanalysable. Theories such as these promise stout ramifications in other philosophical areas, in particular for theories of the mind, of truth, and for the distinctions between the real and the non-real, the objective and the subjective. Although I cannot address this second tier of issues within this work, clearly there is at least the *prospect* that the rule-following considerations have radical and far reaching consequences, and we cannot start to identify these until we know precisely what position we should take with respect to meaning.

As we shall see, I do think that there is a serious problem with our notion of following a rule, indeed one not hitherto identified. The argument is presented in the way it was developed, that is subsequent to a detailed evaluation of existing negative arguments. Of these, there are basically two types - ontological and epistemological - and they are considered in turn in the first two chapters. Neither approach is successful, but it is only in light of their failure - by identifying gaps to be filled, and inspirations to be borrowed - that the new argument could come into being.

As mentioned, *any* argument which threatens the idea that we are rule-followers generates the Kripkean 'paradox' of rule-following. Having presented a negative argument, the discussion opens up: given that there is a difficulty with rule-following, and given that in order to salvage language we have to do *something*, what should we do? In the spirit of investigation I shall not here pre-empt my conclusions any further. I did not set out to defend any one position, and correspondingly I shall allow the argument to take its own course.

I have already indicated one way in which my approach has been influenced by Kripke's enquiry, and there are two further features of his work which I have incorporated here. One is that the rule-following considerations be taken as foundational. Not all interpretations of Wittgenstein accept this view - in particular Wright (1980) sees the discussion of rules as issuing from a strong premise in the philosophy of language, namely the manifestablity of meaning. Again, without deciding which is really Wittgenstein's strategy (my own view is that Wright is here wrong), I shall take the examination of rules as the starting point for our

enquiry, a point from which all other results follow. Our question is what can the examination of rule-following tell us, not what can such an examination tell us in the light of certain other strong philosophical theses.

Secondly, Kripke describes his work, not as a straightforward account of Wittgenstein, but as "Wittgenstein's argument as it struck Kripke" (1982, p. 5). With reference to his own work, Wright, who has probably written more on this issue than anyone else, and whose work has also been a great influence here, could have said something similar. It is this kind of approach which is most exciting: a discussion of rule-following sparked-off, certainly, by Wittgenstein's incredible insights, but one in which exegesis is not allowed to suffocate innovation. It is in the hope of contributing to the discussion in this spirit that I present the following.

PART ONE

AGAINST RULES

1. Kripke's 'Sceptical' Argument

Kripke's (1982) has become the undisputed benchmark for any discussion of the rule-following considerations, and given that it emphasises the negative aims of the initial stages of the argument, it is the obvious place to start our search for a case against the possibility of rule-following. I shall start by giving an account of the core argument - the so-called 'sceptical' challenge - as I understand it. Somewhat surprisingly for an exposition as clear as Kripke's, this argument has been subject to various differing interpretations, and so I go on to say something in defence of the version given here. Finally, I shall consider some of the criticisms which have been directed against the 'sceptical' argument. Although the argument is eventually unsuccessful, our aim here is to assess its strengths and weaknesses, and so to learn what lessons it has to teach, and what improvements have to be made if we are to deliver a more successful argument in its place.

Kripke's argument[1] that rule-following is impossible - that meaning is impossible - proceeds on the basis of exhaustive elimination. He first identifies an adequacy condition which must be satisfied if someone is to be a rule-follower. He then lists all the properties of a person - physical and psychological - which could potentially satisfy this condition. In turn, each type of property on the list is shown to be inadequate in one respect or other. If no property of a person determines that she is following a rule, then she is not a rule-follower. Likewise, since to grasp the meaning of a word is (at least) to grasp a rule governing its correct application,[2] on the basis of the conclusion that nothing determines which rule our speaker is following, then nothing determines what she means by her words, in which case she cannot speak meaningfully.

---

[1] Kripke indicates that he personally does not accept the argument which he presents in the name of Wittgenstein (see for example 1982, pp. 93-94, fn. 76). For the sake of brevity I shall take Kripke's misgivings as read, and speak of the argument *as if* Kripke fully endorsed it.

[2] Of course one cannot often use a single word to utter a sentence, and so the meaning of a word does not usually dictate any correct use outside the context of a sentence. Clearly what is meant is that given the meanings of the other words in a sentence, the word 'green' determines a rule for correct use.

The adequacy condition in question is quite uncontentious, and can be illustrated with an example. Take someone who is following the rule *add 2*. Having proceeded 2, 4, 6, etc., the next number is clearly 8. That is, the rule determines that the correct thing for the agent to do at this stage is to say '8': any other answer would be wrong. Clearly there are an infinite number of steps in the series, and at each one of them the rule must determine what action accords with the rule. Thus the rule determines an infinite correctness condition, an infinite norm. If I am to follow a rule, something must determine which rule I am following, that is, which action is correct in any given situation. This, then, is Kripke's adequacy condition:

To follow a rule, an infinite correctness condition must be in force.

Kripke's 'sceptical' problem concerns *grasp* of a rule (and grasp of meaning), not rule-*following* as such (i.e. with the problem of having an infinite correctness condition somehow in mind, rather than the question of how one uses the rule to inform one's behaviour) - though of course one cannot follow a rule unless one grasps it first.[3]

We should normally think that someone who continues the sequence 2, 4, 6 is following the rule *add 2*, that the next answer they ought to give is 8, and that there is no substantial question as to which rule is being followed. However, in this case the rule *add 2* is not the only option. Kripke makes this plain with his example of the 'quus' function (symbolised by '$\oplus$'), defined as follows:

$$\forall x, y: \; x \oplus y = x + y, \text{ if } x, y < 57$$
$$= 5 \text{ otherwise.}$$

Suppose that we are faced with someone continuing: 2, 4, 6, etc., who has now reached a stage further in the series than she has ever reached before. For convenience we suppose that the speaker has led a rather sheltered life (arithmetically speaking), and so far has had no dealings with any numbers greater than 57. Of course, we should expect her to say that after 56 comes 58, for she appears to be following the rule *add 2*. Yet for numbers less than 57, the series generated by the quus function is (by definition) the same as that generated by addition. So the answers that our subject has given so far are consistent with her following

---

[3] Wittgenstein's attack on the possibility of rule-*following* (i.e. getting into epistemic contact with the requirements of a rule), an attack which assumes that grasp of a rule is unproblematic, is considered in Chapter 2.

either the rule *add 2*, or the rule *quadd 2* (where *quadding* is to quus as adding is to plus). On reaching 56, the two rules diverge: if our subject is adding she ought to say '58' next, if quadding he should say '5'. If she really is following one rule rather than the other, then something about her situation must determine which answer at that stage is correct.

The difference between plus and quus can be treated as a test case. In fact, no matter how many elements in the series have been developed, there are countless such 'quus'-like functions which are consistent with the answers given so far. (We can always define a function which deviates from addition at the very next step not yet considered; and there are endless different such deviations we can consider - add to a certain point, then continue 6, 6, 6, etc., or at first add 2, then add 4, and so on.) Indeed, for *any* response the agent now gives, no matter how the rule has been applied in the past, we are able to construct a rule under which the present answer turns out to be correct. Unless some property of the speaker distinguishes between her following the rule for plus and not quus, then the floodgate is opened; *any* answer is correct according to some rule. So unless something determines that the rule in force is plus, not quus, we should lose sight of the idea that a rule is in force at all.

The situation is not peculiar to mathematics. In a similar vein, we can consider someone who has been using the term 'green' in accordance with normal English usage, namely to refer to green objects. Again, we should usually say that this person means green by 'green', but as before this is not the only option. For example, we can define the predicate 'grue' as follows:[4]

$\forall x, t$: x is grue at time t $\Leftrightarrow$ x is green at t and t < 1st January 2000
x is blue otherwise.

In the (pre-2000) present, those who mean *green* by 'green', and those who means *grue*, ought to apply the term to the same objects, namely the green ones. After this turn of the century we should expect their behaviour to diverge: those who mean *green* ought to continue to call green things 'green', those who mean *grue* ought to apply the term to blue things. For our subject to mean *green*, not *grue*, something must determine that at that time 'green' correctly applies to green objects, and not to blue ones.

---

[4] The predicate 'grue' comes from Goodman (1973). The above definition is significantly different from Goodman's original, but the alteration has become reasonably standard.

The key point to be drawn from these examples is that no matter what actions have been performed in the past, any new action could count as following the same rule, given an appropriate rule. If rule-following is possible, then something about the individual must determine which action is correct in each situation. What we want to know is, what is it about an individual which makes the difference: what determines that one rule is in force and not some other (or no rule at all).

As mentioned, Kripke holds that nothing can determine that some specific rule is in force, rather than some quus-like alternative. To demonstrate this, his strategy is to identify all the properties of an individual which might plausibly play the determining role, and to discount each in turn. The list of putative meaning determining facts he arrives at is as follows:

(i) Self-directed thoughts or instructions
(ii) Dispositions or functional states
(iii) Mental pictures, or other qualitative mental states or experiences
(iv) Platonic objects
(v) Irreducible meaning properties

Kripke argues, in a piecemeal fashion, that none of these things can determine an infinite norm. The intention is that this list covers all reasonable options, that there is nothing else we could suggest as a meaning determinant. On that basis, if none of the above determines meaning, then nothing does.

Self-Directed Instructions

Turning to the first item on the list, we might suppose that the speaker gives herself explicit instructions which would rule out the possibility that she means *quus*. Clearly when we grasp the meaning of addition, we do not (and cannot) think of all possible additions, and so the meaning of the word cannot be determined by our explicit entertainment of all correct applications. However, there are more general principles we might think that our use of the word must obey. For example, our speaker might say to herself "The sum of two (positive) numbers is always greater than either of those numbers." If so, then her instructions rule out the possibility that she means *quus*, for the 'quum' (quus-like sum) is sometimes less than one of the input numbers.

However, this type of instruction is only effective if we can assume that the sentence used to formulate it has itself a definite meaning. In particular, the sceptic will now demand that some fact be produced in virtue of which the speaker means greater by 'greater' and not some quus-like equivalent. We can easily give 'greater' a meaning which is both consistent with the speaker's previous use, and under which the quum of two numbers is indeed 'greater' than either original number. For example, by '<' the speaker might mean 'qreater' (symbolised '⩤'), which is defined as follows:

$$\forall x, y: x \leftthreetimes y \Leftrightarrow \begin{cases} x < y & \text{if } x, y < 57 \\ x \geq y & \text{otherwise.} \end{cases}$$

Indeed, *whatever* instructions the speaker gives herself, the use she has previously made of that expression is finite, and hence consistent with countless different meanings. When the meanings of one's words are in question, it is no use in looking to other words to fix meaning - for this invites us to ask what fixes the meaning of the new expression. If this tactic is maintained, we are left with an infinite regress, forever looking for instructions to fix the meaning of our instructions. Therefore the instruction does not answer the 'sceptic'; it only forces him to shift his attention to a new target.

Dispositions and Functional States

The dispositional thesis which Kripke considers is that meaning *plus*, rather than *quus*, is determined by the responses the speaker is disposed to give. Typically, the suggestion would be that if asked to perform a calculation - for example "What is 67+58?" - she would reply with the sum, not the quum. Similarly, to mean *green* rather than *grue* is to be disposed to apply the word 'green' to green objects. More generally, the thesis is that an object falls within the extension of a term just in case the speaker would apply the term to that object.

Kripke finds this analysis wanting in terms of the infiniteness condition:

> The dispositional theory attempts to avoid the problem of the finiteness of my actual past performance by appealing to a disposition. But in doing so, it ignores an obvious fact: not only my actual performance, but also the totality of my dispositions, is finite. (Kripke 1982, p. 26)

In support of this claim, Kripke notes that addition may involve numbers so large that no human has the intellectual, or for that matter the physical, capacity to deal with them. For example, as numbers get larger, their symbolisations (in standard notation) get longer; eventually numbers get so large that no speaker has the required mental stamina, or a

sufficiently long life span, to deal with them. Clearly the speaker is not disposed to give *any* answer to an addition problem involving such a number, let alone give the addition. As a result, the speaker's dispositions determine a rule only up to a certain point. But unless a unique function is fixed for all numbers, no matter how large, there will be countless quus-like functions consistent with the finite answers the subject is disposed to give, and so what she means is left underdetermined.

In mentioning our inability to perform huge calculations, Kripke identifies but one type of situation in which a speaker fails to accord with the rule being followed, that is in which she makes a mistake. Mistakes can occur for a variety of other reasons - we may misread a figure, forget to carry, and so on. Hence, although someone may means *plus*, she may not be disposed to give the sum of two numbers; in certain situations she may be disposed to give the wrong answer. More generally (i.e. in non-mathematical cases), all sorts of perceptual errors could affect our ability to correctly classify the world around us. For example, under certain lighting conditions, a white plate may look green, in which case we should not be surprised if our speaker calls such a plate 'green'. In this case the speaker is disposed to apply the term 'green' to a white object, but it does not follow that the plate falls within the extension of the term 'green'. Because mistakes are possible, the extension of the term cannot be identified with those objects which the speaker would call 'green'. The simple identification of meaning with one's overall dispositions does not give our words the extensions they actually have, and so does not give a correct account of meaning.

If the extension of a term is to be determined by one's dispositions, then dispositions which produce mistakes must somehow be removed from consideration. In other words, we should focus on how the speaker would act in the absence of any interfering factors which may result in error. To do this, we need specify *ideal* conditions, conditions under which the speaker *is* disposed to perform correctly. The meaning of one's words is not then determined by one's overall dispositions, but only by one's idealised dispositions. These idealisations must cover three areas: *idealised external conditions*, to ensure that the speaker perceives her environment correctly;[5] *idealised internal conditions*, to ensure that, even though the context

---

[5] This condition is more plausible with predicates and terms which apply to material objects, in which case the dispositionalist would require that the lighting, angle of view etc. be in some sense 'normal'. However the point holds even for statements about abstract objects, for, if the speaker is asked a question, she must hear it correctly, she might need pen and paper to carry out a calculation, and she must be able to give her response and perceive it correctly (for example, if she *thinks*, due to an aural illusion, that she said '5' when she meant '125', she may withdraw her answer).

of the utterance is perceived correctly, the speaker does not give a wrong answer (e.g. is not drugged, mentally ill or sheer bloody-minded); and *enhanced mental capacities*, so that she can deal with any calculation, however large. Thus the idealisation takes care of both the need for infiniteness, and for the need to allow for our disposition in certain circumstances to make mistakes.

This third category - idealised mental capacities - faces an immediate difficulty. We suppose that before idealisation our agent was not disposed to give any answer to a particularly large calculation; hence we suggest that the correct answer is the one she would give were she to have extended mental powers, bestowing her the ability to grasp large numbers, and to add them. Any alteration of the speaker's capacity to deal with large numbers will thus *change* her dispositions (this, after all, is the point of the idealisation). However, we could just as well 'idealise' her faculties in such a way that she can grasp large numbers and gives their quum. Either one of these additions is consistent with the meaning determined by her actual (i.e. non-extended) dispositions, since her actual dispositions do not determine any answer in the relevant cases. Given that more than one alteration is possible, how should we decide which alteration counts as 'ideal'? Clearly, her actual dispositions do not determine which idealisation *preserves* what she means, just because her original dispositions *do not determine* what she means. If we are to have a unique, correct idealisation, we have to presuppose that something determines what she means: and therefore we cannot appeal, without circularity, to idealised dispositions to determine what she means.

The same problem arises if idealised dispositions are suggested as a means of accommodating the speaker's disposition to make mistakes in other cases. To return to our earlier example, our speaker applies the word 'green' to green objects when operating under normal (daylight) conditions, but to a white object in abnormal lighting conditions. We might suggest that daylight is ideal, and that the different lighting condition is responsible for the misapplication of the term. Yet, an alternative view would be that the subject actually means *white* by 'green', that it is daylight which produces mistaken applications, and that the supposedly 'abnormal' conditions are really ideal. Although we think that humans can identify colours most successfully in daylight, there is no reason why the reverse should not be true: a being could be more effective is this respect in green lighting than in daylight. So there is no *a priori* reason to say that one type of lighting is ideal, and the other not. Yet this distinction has to be made if the dispositional thesis is to be successful.

The only obvious court of appeal we have is in terms of what the speaker means: if she means green, then daylight is ideal; if she means white, then daylight is not ideal. But this makes the account circular, for we can only say which dispositions are correct given a meaning, and so one's dispositions cannot determine what one means.

The result that ideal conditions can only be specified given a meaning itself points to the fundamental problem which Kripke identifies with the dispositional thesis. For even if we could specify, without circularity, those ideal conditions under which no mistakes are made, the dispositional thesis only tells us what the speaker *would* do, not what she *should* do.[6] In short, dispositions fail to account for the essential normativity of meaning.

The points raised against the dispositional thesis can also be used to discount other naturalistic theories of meaning. The main alternative which Kripke considers is a functional (causal role) theory. As with the dispositional theory, functionalism analyses meaning in terms of input-output, but rather than identifying meaning with the disposition itself, the functionalist states that to mean such-and-such is to be in a state which causes the behaviour in question. For example, to mean *plus* by '+' is to be in a state which causes one to answer addition problems with additions, rather than the disposition itself.

This type of theory offers one immediate benefit, in that it places meaning in an orthodox relationship with behaviour. Normally, we consider grasp of meaning to be a mental state, and that mental states are causally connected to our behaviour. If, for example, we are to explain why the speaker says '125' in terms of the mental states which caused her to utter these words, then one of the states we have to mention is that she means (or understands) *plus* by 'plus'.

Despite this slight advancement, the shift from dispositions to causal roles affords little advantage when it comes to answering Kripke's sceptic. For even though I mean *plus* by '+', I am not *always* caused to give the sum of two numbers - as we know, mistakes are possible. Therefore, as with dispositions, in order to analyse meaning in terms of functional role, we again have to identify ideal conditions - conditions under which I am caused to give the correct answers. Just as before, the only apparent method for doing this is to identify those

---

[6] Cf. Kripke 1982, p. 37.

conditions under which I would accord with my meaning, but in appealing to what I mean, the account becomes circular.

It is not surprising that the move from dispositions to functional states achieves so little. For, whether we say that the speaker is *disposed* to say '125' or is in a state which *causes* her to say '125', these are descriptions of naturalistic states. As such, even if we can specify ideal conditions, we still only have a description of how the speaker would behave, not how she should behave. Naturalistic states - either dispositions or functional roles - do not give rise to semantic normativity.

Mental Pictures and Felt Qualities

The problem of normativity also arises for the next items on Kripke's list - mental pictures, and other qualitative mental states. Taking mental pictures first, the suggestion is that to mean square by 'square' is to have an image of a square before the mind whenever uttering the word. The thought behind the idea is that a picture of a square in some sense refers to an infinite number of squares, in that a picture of a square is (potentially) a picture of each one of an infinite number of different squares in an infinite number of different situations.

The problem with this suggestion is that an image does not determine a norm - there is, for example, no such thing as acting in accordance with an image of a square.[7] Therefore if the image is to play any part in the determination of meaning, it must be suitably *interpreted.* The occurrence of the image must be interpreted to mean that the word associated with the image of a square is to be applied to those objects which *resemble* the image.

Unfortunately, whenever an object stands in need of interpretation, more than one interpretation is available. The resemblance relation mentioned in the above example is only one possibility. Other interpretations we might suggest, which could equally give a meaning to the image of the square, are: apply the word to anything which is the same colour as the mental image; apply the image to anything which is a conic projection of the image; or apply the word to anything which would fit inside the image. It seems we could pick almost any relationship which holds between the image and an external object and use it to construct some interpretation, an interpretation under which the image (or the word associated with the

---

[7] In addition, it is unclear how such an account could deal with the meanings of terms referring to abstract concepts (such as addition, or possibility), which is one of the reasons why Wittgenstein abandoned the picture theory of meaning of the *Tractatus*.

image) applies to the object in question. Yet, if a picture in the mind is to determine an extension for the predicate 'is square', then there must be a *unique correct* interpretation, and something in the speaker's mind must determine which interpretation that is.

What are the characteristics of an interpretation? It is, obviously, just as infinite as a meaning, for it must determine a relation between the (finite) picture and the countless objects the picture represents under the interpretation. The interpretation must also determine a norm - it must determine to what the word associated with the picture correctly applies. Therefore the requirements on an interpretation are precisely the same as the requirements we have identified for meanings - the interpretation must determine an infinite norm. If we have trouble accounting for the infinite norms of meaning, we will have the same trouble accounting for the infinite norms of interpretation. We cannot suggest that the interpretation is another mental image, for then we have another infinite regress. And if we suggest that the interpretation is determined by a *sui generis* mental state, one not to be identified with a mental picture, then we may as well stop the inquiry one stage earlier, and say that meaning is a *sui generis* state. As a result, the appeal to mental pictures adds nothing to the account of meaning, but just defers our difficulty.

In terms of meaning determination, distinctive felt qualities are just as unsatisfactory as mental images. To take an example, a headache does not in itself determine that any action is correct; there is no such thing as acting in accordance with a headache. If a felt quality is to determine a norm, then the felt quality must be interpreted in an appropriate manner. Again, the question is: what determines that a certain interpretation is correct? If it is another qualitative state then the account is regressive - for that qualitative state will require interpretation in order to fix the correct interpretation of the initial qualitative state. If, on the other hand, the correct interpretation is determined by some property other than a felt quality, then that property must be capable of determining a norm without the contribution of the felt quality. Once again, the account is either regressive, or the object which requires interpretation is superfluous, which is why felt qualities can play no part in meaning determination.

Platonic Entities and Irreducible Meaning Properties

The remaining two suggestions may be treated together. The first is that the meaning-determining state is not reducible to dispositions, or qualitative mental states, but is rather a *sui generis* mental state. The other is that meaning is determined by the fact that one's mind

grasps a Platonic object - an abstract object which determines an extension as a matter of essence. With respect to the *sui generis* hypothesis, Kripke complains that this manoeuvre is unexplanatory:

> Such a move may in a sense be irrefutable, and if it is taken in an appropriate way, Wittgenstein may even accept it. But it seems desperate: it leaves the nature of this postulated primitive state...completely mysterious. (Kripke 1982, p. 51)

Moreover, there is a distinct logical difficulty with this answer, for "Such a state would have to be a finite object, contained in our finite minds" (p. 52). How such a finite state can issue in an infinite norm is, Kripke suggests, beyond comprehension:

> Can we conceive of a finite state which *could* not be interpreted in a quus-like way? How could that be? (p. 52)

> it remains mysterious exactly how the existence of any finite past state of my mind could entail that...I must give a determinate answer to an arbitrarily large addition problem. (p. 53)

The argument is this. If a finite state is to yield an extension over an infinite domain, then the state must (surely?) be interpreted in some way - for how else can the answers to infinitely many question be determined by a finite object? And if the meaning-state has to be interpreted, something in the speaker's mind must determine what the correct interpretation is. Which is to say that another finite object must determine an infinite norm, and so it in turn must be interpreted. (Once again the regress of interpretations.) This regress is clearly vicious, for the speaker's finite mind cannot possibly contain an infinite number of interpretations. Hence a *sui generis* state cannot determine what the speaker means.

The theory referring to Platonic objects succumbs (Kripke claims) to a similar objection. The supposed advantage of this position is that there is no difficulty with the idea that an abstract object can determine an infinite extension from its own resources, as it were - that is, without interpretation. Yet to govern my behaviour, such an object must be grasped by my mind (in Frege's terminology, I have an idea in my mind which grasps the Platonic sense), and here again we have a difficulty in reconciling the 'finiteness' of mind with the infiniteness of extension. As Kripke says:

> The idea in my mind is a finite object: can it not be interpreted as determining a quus function, rather than a plus function? (Kripke 1982, p. 54)

As before, a finite object can be interpreted in many different ways to yield different infinite extensions. This, Kripke claims, shows that no such finite state can fix a unique extension.

The importance of this part of the argument should not be underrated, for without an argument against irreducible meanings, the fact that Kripke fails to find a meaning-determinant is simply proof of the hopelessness of a reductionist programme. As with any other property, there is no automatic requirement that we should be able to provide an analysis of grasp of meaning, to say what meaning consists in any other terms. Should it turn out that we cannot give any account of a meaning determining fact, if we cannot analyse the notion in some other terms, we do not thereby have any right to say that no such property exists, or remains uninstantiated.[8] Success in eliminating this type of response is therefore vital for the 'sceptical' argument to succeed. Unfortunately, as we shall see, Kripke's argument is at its weakest here. In particular the basis for the claim that an infinite norm cannot be contained within a 'finite' mind is not particularly robust. To address this issue takes us away from exposition into criticism of Kripke's argument, and so I shall for the moment defer further discussion.

Concluding the 'Sceptical' Argument

This brings us to the end of the list of putative meaning determining facts. If Kripke has shown that every item on his list is inadequate for the determination of meaning, and if we accept that his process of exhaustive elimination has indeed been exhaustive, then he has shown that meaning is impossible. We think that if I start reciting the series 2, 4, 6..., and state that I am at every stage adding two, then every element of the series is determined in advance, that at each stage in the sequence there is one correct answer, and that this answer is determined by the meaning of the word 'add'. According to the 'sceptical' argument this is false: nothing determines that any answer is correct or incorrect, and no answer is any better than any other. Quite clearly the thesis is quite general in its application: with *any* word, no matter how convinced we are in our own practice that there is a correct standard of use, the result of the 'sceptical' argument is that this is an illusion, putting language as a whole on the verge of collapse.[9]

---

[8] Several authors have criticised Kripke for unwarrantedly adopting a reductionist position. For example, Goldfarb states that "Kripke's concern is with physicalist reductions of meaning notions" (1985, p. 479). The fact that Kripke considers meanings as *sui generis* states shows that he does not take this dogmatic reductionist line.

[9] I do not intend to give a detailed account of the differences between Kripke's account of Wittgenstein's negative argument and Wittgenstein's actual argument. Certainly much of what Kripke says is a faithful reading (rather more faithful than many critics have suggested). For example the following elements of Kripke's exposition are explicit in the *Investigations*: the concern with what constitutes meaning and understanding (e.g. PI §148, §153, and §186); the claim that no qualitative state could constitute meaning (most clearly stated in the discussion of reading PI §156-178); also that pictures and self-directed instructions fall to the regress of interpretations (e.g. PI

Some Presentational Issues

As I mentioned at the beginning of this chapter, the above exposition of the 'sceptical' argument is intended to give the core thesis as straightforwardly as possible, and as such it has been stripped of some of the presentational devices which Kripke employs. The reason for this is that some of the superficial elements of Kripke's original exposition are misleading - indeed, they have led various early commentators to misrepresent the nature of the 'sceptical' argument. Whilst the early discussion provoked by Kripke's book has, I think, diagnosed the sources of error - and settled that such misrepresentations are indeed misrepresentations - it is important that the relevant issues be identified, not least to establish that the elements I have omitted are indeed wholly dispensable. Once these are out of the way we will be in a position to assess some rather more substantial objections.

The first observation is that although Kripke describes his argument as a "sceptical" argument (and his protagonist as "the sceptic"), this label is misleading. Kripke's argument has nothing to do with classical scepticism, nor indeed with any epistemological issue. This point is important because if we were misled into thinking that the 'sceptical' argument really were sceptical, then we would perhaps dismiss it rather too quickly. Sceptical doubts - real sceptical doubts, doubts about the possibility of a certain kind of knowledge - do not in general allow us to make conclusions about the way of the world. Usually the fact that we do not *know* that p does not entitle us to conclude that not p. Hence, if all Kripke shows us is the sceptical claim that we do not *know* what a speaker means (or perhaps that she does not know what she herself means), we cannot thereby infer that she does not mean anything at all. Thus, as a means of establishing the impossibility of meaning, a sceptical argument looks unpromising.[10]

---

§139, §§189-90); dispositional and functional reductions fail (e.g. PI §149 and §§193-195 - though Kripke's discussion is a considerable advancement on Wittgenstein's, especially with respect to dispositions), and mysteriousness of how a rule can determine all the steps in advance (e.g. PI §188). As we shall see in Chapter 2, Wittgenstein had additional epistemological concerns (which Kripke, I think wrongly, interprets as yielding an ontological point when taken to their limit). In addition, there are doubts over whether the *aim* of the investigation is quite the same as Kripke makes out, or whether the end result Kripke offers has anything acceptable to Wittgenstein. My own view is that Kripke's discussion is rather closer to Wittgenstein than many have made out in terms of both negative argument (not least because many such critics mistakenly identify the 'sceptical' argument as a genuine kind of scepticism, something Wittgenstein would certainly abhor), and also in terms of the positive proposal.

[10] The mistake of ascribing an epistemological argument to Kripke is made by Baker and Hacker (1984). McGinn makes a similar error (1984, p. 72). I shall continue to follow Kripke's (now established) usage of the term 'sceptical', but shall continue to signal the inappropriateness of the nomenclature with scare-quotes.

The problem is not limited to Kripke's use of the term 'sceptical', for he does portray the argument as epistemological by talking in terms of justification. For example, he says:

> In the discussion below the challenge posed by the sceptic takes two forms. First, he questions whether there is any *fact* that I meant plus, not quus, that will answer his sceptical challenge. Second, he questions whether I have *any reason* [emphasis added] to be so confident that I should answer '125' rather than '5'....An answer to the sceptic must satisfy two conditions. First, it must give an account of what fact it is...that constitutes my meaning plus, not quus. But further, there is a condition that any putative candidate for such a fact must satisfy. It must, in some sense, *show* [emphasis added] how I am justified in giving the answer '125' to '68+57'. (Kripke 1982, p. 11)

The first condition mentioned is familiar - the meaning-determining fact (property) must determine a unique function over an infinite domain, and it must do so normatively. However, the second condition is new: Kripke here claims that the speaker himself must be able *to tell* that his action is in accordance with his previous meaning if he is to be fully justified in acting as he does.[11] As a result, Kripke's argument might be read as follows: if the speaker cannot *verify* what she meant in the past (in order to decide how to act now), or cannot *justify* the claim that her meaning requires her to say '125' (rather than '5'), then there is no fact about what she means; none of the items on the list provide such *justifications*, so there is no fact about what the speaker means. Of course the fact that a speaker cannot justify her claim to mean *plus* rather than *quus*, or that she ought to answer '5' does not mean that these claims cannot be true. Any argument which followed this course without additional material would clearly be verificationist, and Kripke for one gives us no reason, in this context, to accept such a move.

As it happens, though, Kripke does not reject any of the potentially meaning-determining properties from his list on the basis that it cannot provide the justification mentioned above. Therefore, in the actual execution of the 'sceptical argument', the condition that the meaning determining fact must provide such a justification is redundant. Indeed, Kripke explicitly states as much:

> it is clear that the sceptical challenge is not really an epistemological one. It purports to show that nothing in my mental history or past behaviour - not even what an omniscient God would know - could establish whether I meant plus or quus. (Kripke 1982, p. 21)

As we shall see, the search for a justification for following a rule in the way one does is one of Wittgenstein's central concerns, and it is therefore a flaw in Kripke's interpretation of Wittgenstein's argument that he (Kripke) makes no use of it. I shall consider the

---

[11] In fact Kripke introduces the sceptic as one who "questions my certainty about my answer" (1982, p. 8).

epistemology of rule-following in Chapter 2, but for now I merely record the stance to take with respect to Kripke: his presentation suggests epistemological concerns, but the substance of the argument does not bear this out.[12]

The second clarification to make is that I have portrayed the relationship between meaning and action as a synchronic relationship - my present meaning determines how I should act in the present - whereas Kripke describes it as a diachronic relationship. For example, he says:

> [The sceptic] questions whether my present usage agrees with my past usage, whether I am *presently* conforming to my *previous* linguistic intentions. (Kripke's emphasis, Kripke 1982, p. 12)

According to this passage, Kripke's search is for some property of his past self which determined what he meant in the past; his *present* intention is to act in accordance with that *past* meaning, and so his *past* meaning determines what he ought to do *now*. Thus the normativity of meaning is something which apparently stretches from the past into the present.

If Kripke's argument rests on a diachronic notion of semantic normativity, then it rests on a mistake. Meanings are not normative over time. For example, it may well be that the speaker meant *plus* in the past, but from this fact alone it does not follow that she should, in the present, use '+' in accordance with that prior meaning; nor, for that matter, does it follow that she should continue to mean *plus* by '+' in the present. What she means now is independent of what she meant in the past, and the present intention to conform to a previous meaning is not essential for present meaning. What the speaker meant in the past determined what she ought to have said in the past; what she means in the present determines what she ought to say in the present: meanings are normative only *at* a time, not *over* time.

Despite appearances, Kripke does *not* make this blunder about the temporality of semantic normativity. This talk of the ability of a linguistic norm to stretch into the future arises because Kripke considers a speaker who intends to accord with his previous meaning.[13] If you intend to accord with a previous meaning, then clearly that meaning does set a standard

---

[12] At times Kripke appears to conflate that which justifies the speaker saying '125' and that which makes '125' the correct answer. For example, he says (1982, p. 37) that dispositions cannot account for normativity, and then says that dispositions provide no justification for any action. As I read it, the latter is intended to be a re-iteration of the former point, meaning that Kripke identifies that which makes the judgement true with that which justifies the judgement.

[13] See for example Kripke (1982, p. 8).

for present use. But, as stated, that intention is not essential to meaning, and so is not a *semantic* norm.[14] Nevertheless, the possibility of intending to accord with previous meaning does highlight the fact that meaning is a normative notion. When Kripke concludes that nothing about his past self determines what would count as satisfaction of his present intention (to accord with that past meaning), he simply makes the claim that nothing in the past determined what he meant in the past. As he goes on to say (Kripke 1982, p. 13), if there is no fact about what he meant in the past, then there is no fact about present meaning either, just because there is no relevant difference between the past and the present which could render meaning impossible then, but possible now. It is this failure of meaning - at any one time - which is the proper result of the 'sceptical' argument, not the inability to accord with past meaning.

The 'temporal spread' idea is thus not germane to Kripke's argument, but is a device which is *supposed* to make the argument clearer. The reason for stating the problem in these terms is that Kripke presents the 'sceptical' challenge from first-person perspective; and thus Kripke's own words constitute both the meta-language and the object language.[15] Since the distinction between the object-language and the meta-language is all-important for his argument, it would be highly damaging if they were conflated. Keeping them temporally separated (past/present) is simply a means of keeping them conceptually separated.[16] Consequently nothing is lost, and clarity is gained, by presenting the problem in terms of

---

[14] This mistake arises from a confusion between the normativity of meaning and the normativity of intention. Of course, if the speaker intends to mean the same thing as before, there is something she ought to do, namely mean the same as she meant before (glossing over the doubtful implicit claim that 'meaning something' is something we do). Thus intentions are akin to norms, in that intentions determine satisfaction-conditions. However meaning does not rest on the notion of intention (or at least not in this sense) - the speaker does not have to intend to mean the same in order to mean the same. The norm of this intention is therefore not a semantic norm.

[15] Kripke presents the argument in terms of the first-person in order to show that no behaviouristic restrictions should be assumed to operate on the type of property we might cite as meaning-determining (see for example Kripke 1982, pp. 55-56). Indeed, Kripke differentiates sharply between his 'sceptical' problem and Quine's argument for the indeterminacy of translation (which is certainly based on a behaviourist premise - see Quine 1992, p.37). Kripke, unlike Quine, thinks that appeal to introspective mental states is perfectly legitimate in this context, and may be utilised in the hope of refuting the 'sceptic'. His use of the first-person highlights this fact (Kripke 1982, p. 15).

[16] Kripke explains that he talks about past meaning in order to make object language and meta-language easily distinguishable (1982, pp. 12-13). Nevertheless, his use of a temporal spread has led to genuine confusion. For example Colin McGinn (1984, p. 174) takes normativity over time to be essential to the argument. (McGinn's error is pointed out by Boghossian (1989, pp. 90-91).) In addition, Coates (1986) and Sartorelli (1991) both place undue weight on the trans-temporal aspect of the argument. Kripke's statement that "The relation of meaning and intention to *future* action [emphasis added] is *normative*, not *descriptive*." (1982, p. 37) only adds to the impression that trans-temporality is a vital factor, whereas in reality it is not.

present accord with *present* meaning, as I have done. In my view, the meta-language/object-language distinction is more clearly made by considering the situation of a third person.

Objections

With these presentational issues out of the way, we can turn to the more substantial objections which have been raised against the 'sceptical' argument. These fall into two main categories:

(a) Kripke's dismissal one or other of the items on his list is unsuccessful.

(b) Kripke's list is not exhaustive: some fact which could be meaning-determining is not considered.

Little argument has to be given to establish the second point, that Kripke's list is not fully exhaustive. For example, Kripke does not consider the possibility of a causal theory of reference (as pointed out by Goldfarb 1985; McGinn 1984a; and Maddy 1984); nor that meaning might consists in a capacity (cf. McGinn 1984a, pp. 168-175); nor that meaning may be analysed along Gricean lines -that is in terms of the propositional attitudes (beliefs, intentions, and so forth) which accompany the utterance of a word (cf. McGinn 1984a, pp. 167 -168).[17] I do not mean to suggest that any one of these properties is a particularly strong candidate, but Kripke does indeed fail to discount them explicitly.

This type of objection, though, ought to take second place in our order of priorities. For, to pursue this line of criticism, we should have to discern whether any one of these freshly suggested properties can be meaning-determining; and to do that, we should have to show that the arguments Kripke uses to discount those items which are on his list cannot be applied with equal efficacy against those which he fails to consider. So, before suggesting that the 'sceptical' argument fails because Kripke's list is not exhaustive, we ought first to assess how effective the objections raised against those items explicitly mentioned on the list actually are. Only once we have the exact measure of Kripke's resources can we see if there is any property, on or off the list, which is capable of surviving all the objections raised. It is therefore prudent to postpone the examination of what Kripke fails to consider until the success (or otherwise) of what he does give us has been properly gauged.

---

[17] The classic works advocating this type of theory are Grice (1957; 1969), and Searle (1969; 1979). Although McGinn mentions that Kripke fails to discount such a theory as providing a meaning-determinant, he does not endorse this as an adequate response to the 'sceptical' argument (cf. 1984, p. 168).

Of those items on Kripke's list, several do not require further examination. The arguments given that meaning cannot be determined by self-directed instructions, mental pictures, or felt qualities are incontrovertible. Any property which is meaning-determining must do so without the need for interpretation, and so these entities can be safely removed from consideration. The results of the other aspects of the argument remain less secure, and these are the focus for the rest of this chapter. The two contentious elements of the 'sceptical' argument are those which relate to Kripke's discussion of a dispositional theory of meaning, and those which relate to Kripke's objections to the *sui generis* meaning property and Platonic object theories. I shall consider each in turn.

The Dispositional Thesis

As we have seen, Kripke claims that meaning cannot be dispositional because meanings are both infinite and normative, whereas dispositions are neither. In order to reinvigorate the dispositional thesis, respondents to Kripke have taken these objections head on, arguing that dispositions are indeed infinite in the required sense, and can yield normativity.[18] The present task is to assess whether Kripke's critics are successful in establishing these counterclaims.

Dispositions and Infinitude

It will be recalled that the problem of infiniteness is, in the terms of our arithmetical example, that numbers eventually get so large, and their symbolisations so lengthy, that the human mind is incapable of dealing with them. No one could mentally grasp such numbers, nor live long enough to fully survey them, and so no one actually has the disposition to give the sum of such numbers (nor, indeed, the quum). Consequently, human dispositions cannot determine that '+' means *plus*, for plus is defined over the infinite domain of natural numbers.

---

[18] Some of the many criticisms that have been directed against Kripke's rejection of a dispositional reduction are off-target. For example, Coates (1986) and McGinn (1984a) both succumb somewhat to the confusion arising over the notion of trans-temporal normativity mentioned in the previous chapter; consequently they do not take the notion of synchronic normativity seriously, and falsely conclude that dispositions do not have to account for this type of normativity. In turn, Philip Pettit (1990a) thinks that the job of the dispositionalist is not to give a constitutive account of rule-following, and hence of meaning, but to save the *phenomenology* of rule-following. At least, in giving a dispositionalist response to Kripke, Pettit describes his project as the "attempt to give an explanation of how a rule-follower may *see* herself as having made a mistake and an explanation therefore of how we may *see* her inclination as having misfired." (Emphasis added, Pettit 1990a, p. 16.) But this is certainly off the point, for in order to answer the 'sceptic', we do not need an explanation of how we *appear* to make mistakes, but of how anything could actually *be* a mistake.

Kripke suggests that the way to solve this problem is by suggesting certain *idealisations* to our subject - to augment the speaker's cognitive capacities in such a way that she could grasp huge numbers, and also perhaps scan them at an accelerated rate. Then we could claim that the speaker's *extended* dispositions determine the answers to calculations over the whole range of integers, for the speaker's dispositions *are* infinite when suitably idealised.

Kripke's objection to this suggestion is that any such augmentation of the speaker's dispositions gives rise to a vicious circularity. If we are to alter the speaker's capacities then there is clearly more than one way in which this could be done. For instance we could change her brain in such a way that she is disposed to compute sums; alternatively, we could propose an alteration that would leave her disposed to compute a quus-like function. Both 'idealisations' would preserve the dispositions the speaker has with respect to smaller calculations. Yet we want to know which idealisation accords with what the speaker means. Why should we choose one idealisation over any other? The only grounds we could possibly have for preferring one cognitive extension over the other is that only one endows the speaker with the correct dispositions - correct, that is, *given* what she means by '+'. The only standard we could possibly have for preferring one cognitive alteration over another is provided by the meaning of the word, and so the theory presupposes the very thing it is supposed to describe. It is this circularity which makes the specification of augmented capacities impossible.

This argument is rebutted by observations made by Blackburn (1984b),[19] which concern the nature of paradigmatic dispositions. To take an example of Blackburn's, the fragility of glass is a perfectly ordinary disposition, and one which is readily seen to be, in one respect at least, infinite. That is, a glass will shatter in any number of different ways, when struck in any number of different places, in any number of different locations, by any number of different objects. The disposition thus yields an 'output' (the glass shattering) for an infinite number of different 'inputs' (the glass being stuck with a hammer on Earth, with a brick on the moon, and so on), and it is in this sense that the disposition is indeed infinite.

On its own, this observation is not sufficient to overturn Kripke's charge that dispositions are finite, for it is not obvious we are talking about the right kind of infiniteness. One important

---

[19] Blackburn's criticism of Kripke is echoed by Forbes (1984, pp. 233-235) and Ginet (1992, pp. 67-71).

sense in which extensions are infinite is that words refer to objects throughout (infinite) time and space ('green' refers to green objects no matter how far away temporally or spatially). And the point made above, that dispositions account for an infinite number of different inputs in a *finite* portion of the universe, does not address this issue at all.

The reason why it might be thought that dispositions fail of this kind of infiniteness - that they do not 'extend' throughout time and space - is that physical objects cannot reach faraway locations without undergoing certain changes, which means changes to their dispositions. To continue with Blackburn's example, glass has a certain degree of structural instability; in the time it would take to travel an immense distance the glass would disintegrate. Hence it is not true that the glass would shatter if struck in a distant region of the universe. Correspondingly, a speaker would not call a faraway galaxy 'green' (at least not on the basis of empirical evidence) for she would die before she could ever be transported into a position from which she could see it. In that case, on the simple dispositional analysis, this green galaxy would not fall under the extension of the speaker's term 'green', making the dispositional theory inadequate.

In fact this objection - that the glass is not disposed to break on Alpha Centauri, and the like - rests on a misunderstanding. When attributing the glass with the disposition to break we are not claiming that the glass would be disposed to break *after* undergoing a million-year journey. We are concerned with the dispositions the glass has now; in other words how the glass in its *present state* would behave under *different* circumstances. The fact that Alpha Centauri is so distant does not have any bearing on the question of what would happen to a given glass if it were in the same state *as it is now* if it were in this different location. The same can be said with respect to time. When asking how a glass would behave in the year 5098, we are not interested in how a 4000 year-old glass would behave; but rather, how a glass *as it is now* would behave if it were to exist at this different temporal location. In both cases it is perfectly legitimate to ignore any processes we might *actually* have to put the glass through in order get it into the appropriate context, for such processes do not play any role in determining the truth or falsity of the (counterfactual or subjunctive) conditional in question. So when properly characterised, there is no problem in attributing dispositions to objects throughout space and time.

These results, which concern the dispositions of objects, can readily be applied to people, and their linguistic dispositions. Just as glass is fragile throughout time and space, we may

conclude that we have the disposition to apply 'green' throughout time and space, even though we could not actually reach many distant locations. For predicates referring to empirical states of affairs, there is no problem arising from considerations of infiniteness.

Blackburn suggests that the foregoing considerations can successfully be transferred to the mathematical case, even if it does require a slightly more sensitive treatment. As a first step he suggests that we consider, not the response the speaker is disposed to give to an arithmetical problem, but the answer he is disposed to *accept*. The answer he is disposed to accept is "the one that would be given by reiterating procedures I am disposed to use" (1984b, p. 289). The procedures in question are the subroutines we use to add large numbers, such as adding columns of single digits, carrying tens, and so on. So whereas my disposition to carry tens will at some point give out, there is still some answer I would give if I were to continue carrying ten, which is why the answer I would accept offers an advantage over a mere disposition.

At first blush, it may look as though Blackburn appeals to one rule (the rule for reiteration) in order to establish another (the rule for addition), in which case his account would beg the question. More specifically, the speaker's manifest behaviour in any finite surveyable sample, in which she carries tens, is consistent with both the claim that she is carrying ten, and with the claim that she is 'quarrying' ten (carry ten for the first thousand digits, then carry twenty). In order to say that there is a determinate answer she would accept, something must determine which process it is which is to be reiterated, which under the dispositional theory can only be the speaker's dispositions. In claiming that there is a determinate process which can be reiterated, Blackburn claims that her dispositions determine what she ought to do in every situation, for every number. Thus the disposition to accept certain answers only determines an infinite extension by appeal to a disposition which determines an infinite extension, which is precisely the issue in question.

At this point the observations made about dispositions to behaviour applying throughout time and space have some bearing. There is no doubt I would not be disposed to reiterate certain arithmetical procedures indefinitely, simply because I would get bored, or die, first. But in saying that I would die first, we make a comment not about what I would say were I in the position to assess the matter, but only what my disposition would be *after* undergoing the process necessary to get me into that position. The objection that I am not disposed to carry ten is, as Blackburn suggests, akin to the claim that the glass cannot get to Alpha

Centauri, which, as we have seen, is not the basis for a strong objection. If any disposition is to be meaning-determining it must be what I would say were I in the position to assess the matter, not the disposition I would have after undergoing a process which alters my dispositions. The fact that I would die in getting into the required position does nothing to indicate that there is not a determinate answer I would give were I in that position. Consequently there is no difficulty in saying that I am disposed to carry, no matter how great the numbers involved, and so no reason to suppose that I am not disposed to give additions for all pairs of natural numbers.

The above points fully rebut Kripke's argument concerning infiniteness, but it should also be noted that there is actually no need to respond to Kripke's charges directly. It is entirely possible for the dispositionalist to fully accept Kripke's charges, for the claim that the finiteness of dispositions renders them incapable of determining meaning is something of an overstatement of the situation. The dispositionalist need only concede to the sceptic that dispositions are not adequate for the determination of functions over certain infinite domains, and that, boldly, the function identified by our word 'plus' is not determined over all natural numbers. Whilst Kripke's argument is that such a result would be wholly destructive of the notion of meaning, in reality the point is not so forceful.

The claim that dispositions determine what the speaker means, together with the claim that her dispositions are finite, does not entail that the speaker's words are meaningless. At best it shows that what she means is, to a certain extent, indeterminate: nothing determines the answer she ought to give for sums above a certain threshold. If the threshold is very low, then certainly the indeterminacy is too great to bear: there would be nothing present which we could possibly call 'meaning'. If, however, it is only after a few billion or so that the correct use of '+' is left undetermined, there is nothing to prevent the claim that '+' has some kind of content. In other words, the notion of meaning could withstand this level of indeterminacy. If it transpired that this was the only reason the sceptic could find for rejecting a dispositional account of meaning then his position would be weak, to say the least.

Normativity

The more significant objection to the dispositional theory is that dispositions cannot account for the normativity of meaning. As mentioned above, Kripke's argument comes in two parts. The first is that the dispositional theory cannot accommodate the possibility of error: I mean

red by 'red', but I am not disposed to call all and only red things 'red'. A satisfactory dispositional theory must therefore be buttressed with the addition of *ceteris paribus* clauses, which would remove all sources of error. (What I mean is then identified by the answers I am disposed to give under these conditions.) Kripke objects, as we have seen, that such *ceteris paribus* clauses can only be supplied *given* what I mean, the theory thus being viciously circular.

The second objection is that even if I am disposed to call all green things green, the dispositional theory would only state what I *would* say, not what I *should* say. The point is that there is a fundamental difference between meanings and dispositions: meanings are normative, dispositions are not.

Before addressing these issues directly, it is important to be clear about what is involved in semantic normativity. So far I have presented the matter in the terminology used by Kripke, that is as a distinction between what one would do, and what one should do (for examples of Kripke's use of such language see his 1982, p. 24, p. 29 p. 37 and p. 56). In stating that a meaning determines what one *ought* to say, Kripke appears to hold that a meaning determines an absolute (or categorical) norm: that one ought to speak the truth *no matter what*. But clearly this is not so - I *ought* call red things 'red' only given that I intend to speak the truth. If I intend to lie, then I ought to call red things anything but 'red'. That is to say, the force of the 'ought' of meaning is provided by the generally prevalent intention to speak the truth, and likewise a rule prescribes what we ought to do only given the intention to follow it. The 'ought' of rules and meaning depends on one's aims, and so is not categorical, but hypothetical.

Importantly, though, this dependency of the 'ought' on one's intentions does not mean that the normativity of meaning is provided solely by the intention to speak the truth. (If that were the case, then the problem of semantic normativity would reduce to that of accounting for intentional mental states.) On the contrary, the intention to speak the truth can only be satisfied on condition that something counts as speaking the truth; the intention to follow a rule can only be satisfied if some action counts as acting in accordance with it. So to speak the truth, or to follow a rule, there must be a correctness condition of the appropriate kind which is quite independent of one's intentions. It is the notion of correctness which makes meaning a normative notion, and so the challenge for the dispositionalist is not to get an 'ought' from a 'would', but rather to construct a notion of correctness.

Despite the slight lack of clarity Kripke brings to the characterisation of semantic normativity, this in no way invalidates his argument, and the points he raises remain wholly pertinent. The various aspects of Kripke's argument at this point may be drawn together by putting the matter on a slightly more formal footing. As stated, the aim is to provide a set of ideal conditions under which any subject will give only correct answers. That is, we need some set of conditions O which satisfy the following conditional:

$\forall$S, x: Under O, S would apply 'F' to x $\Leftrightarrow$ 'F' is true of x.

If we could provide a condition O which satisfied comparable conditionals for all terms in English, then we should have shown that any utterance made under O conditions is true, and that an utterance is only true if it is what would be said under O conditions. If this could be generalised, we would have shown that:

$\forall$s: s is true $\Leftrightarrow$ s would be uttered under conditions O

(where s ranges over sentences). In that case, since what is true is necessarily co-extensive with what one is ideally disposed to say, then there is no obvious reason not to reduce one to the other, so that ideal dispositions would yield a correctness condition.[20]

To rebut Kripke's objections to the dispositional thesis, two questions have to be addressed. The first is whether this is really all there is to semantic normativity. The second is whether there are any such O conditions. The first question is difficult to answer, for there are competing intuitions. On the one hand, the truth of the biconditional above would ordinarily be taken as excellent grounds for a reduction. On the other hand, attendant with any such analysis of truth is an implicit claim about the relevant order of determination. That is, the claim is not only that true utterances are co-extensive with idealised responses, but in addition that such utterances are true precisely *because* they are made under ideal conditions. Whilst not immediately objectionable, this result is uncomfortable, for as Wright notes (1987, p. 393), this makes the correct application of a rule - and hence a concept - dependent on the responses of the subject. And if every subject matter is response-dependent, then we

---

[20] Forbes (1984, p. 228-232), Goldfarb (1985, pp. 477- 478), Ginet (1992, p. 70), and Millikan (1990, p. 337) all advocate, in direct response to Kripke, the thesis that the specification of ideal conditions is sufficient for normativity, and that such can be given without circularity.

have reached some form of global subjectivism - not something the (presumably realist) dispositionalist will have bargained for.

No doubt more has to be said if Wright's objection is to be made fully convincing (and for my part I am not sure that the argument can be secured), but I shall not pursue such matters, for the point is that although there are difficulties the dispositionalist has to overcome - and although the result might not be quite as intended - the burden of proof is not to establish the dispositional thesis, but to refute it; and so far no concrete refutation has been forthcoming.

The remaining question is whether Kripke's contention that there can be no such O conditions is any more convincing.[21] As we have seen, in the first instance Kripke suggests that ideal conditions can only be specified if we are given what the subject means. Putting the example used earlier into the present terminology, suppose that there is a set of ideal conditions O (perhaps including the specification of daylight) under which I am disposed to call all and only green things "green", so that we have:

$\forall$x: Under O, I would apply "green" to x $\Leftrightarrow$ x is green.

Suppose further that under conditions Ø (perhaps including the specification of green lighting) I am disposed to call all and only white things "green", so that we have:

$\forall$x: Under Ø, I would apply "green" to x $\Leftrightarrow$ x is white.

---

[21] To counter the claim that ideal conditions cannot be specified, Forbes (1984) suggests that no such specification is required, for any dispositional property carries with it *implicit* ideal conditions. For example, glass will not break if the temperature is very high, salt will dissolve in water but not petrol, and just about anything will burn if powdered. Yet these facts do not entail that that glass is not fragile, that salt is not soluble, or that aluminium is inflammable. By analogy, if we identify the speaker meaning *plus* with her disposition to use the word 'plus', we *already* allow for the fact that in some circumstances she will behave in a non-ideal manner. It does not matter that we cannot spell out what the ideal conditions are without circularity, for we do not have to spell them out.

Yet the ideal conditions which Forbes mentions merely record conditions which human beings feel entitled to take for granted. It is only because water is so abundant that we take soluble to mean soluble in water (chemists do not - they always specify the agents involved), and because many dispositions do not vary with the few degrees Celsius that air temperatures vary by the effect of temperature are usually ignored. Thus the ideal conditions Forbes talks of are not objective in any sense but merely record the background conditions which humans have come to expect. With respect to meaning there are no commonly accepted ideal conditions, for different perturbing factors will affect the application of some concepts, but not others. It is, therefore, not possible to appeal to the notion of implicit ideal conditions to solve the problem.

On the face of it, we can only say that conditions O are ideal, and that Ø are not, on the basis that I mean *green* by "green", which begs the question.

However, it ought not be surprising that, when presented with a set of putative ideal conditions in this arbitrary and piecemeal way, we have no means of making a decision. Whilst in the normal case we recognise colours better in daylight than, say, twilight, there is no reason why this situation could not be reversed, and that for some speakers twilight is ideal, daylight not. So without further information, we cannot hope to legislate on the matter. Indeed, if we are to provide an analysis of truth, we need to specify ideal conditions in a way which applies quite universally, to all speakers, and for all concepts, which is to say that we need a *principle* governing whether any given situation is ideal or not. The fact that we cannot decide between explicitly specified conditions for a particular individual shows only that we have not yet identified the underlying principle, but does nothing to establish that no such universal specification is possible.

Kripke does present an auxiliary argument to support his point. He rejects the possibility of any specification of ideal conditions for the additional reason that an error may occur at any time even *in the absence* of perturbing factors. To show this, Kripke considers an example in which:

> the subject has a *systematic* disposition to forget to carry in certain circumstances: he tends to give a uniformly erroneous answer when well rested, in a pleasant environment, free from clutter etc. (Original emphasis. Kripke 1982, pp. 31-32)

Kripke here stresses that the 'error' is not the result of extraneous environmental factors, nor of physiologically based defects in the subject - he simply, yet persistently, forgets to carry. The same dispositions are therefore consistent with two meanings - in the one case the failure to carry is a mistake (under the assumption that the subject is adding), under the other this oversight is not a mistake at all, for the subject is calculating some other function which does not require carrying. Either option is compatible with the subject's overall dispositions. Kripke claims that since it is possible for two people to share the same dispositions yet to mean different things, there can be no one set of ideal conditions which yields 'the' extension fixed by the common disposition.

Unfortunately, it is not wholly perspicuous that, in the absence of any disturbing factor, the type of systematic error that Kripke's example trades on is possible. For in the given situation, if it were pointed out to the subject that, by the normal standards of addition, he

had made an error, and at which point in the calculation the 'mistake' was made, then if he really does mean *plus*, we should expect him to revise his opinion. If, on the other hand, he stuck to the original verdict, we should have to accept that he means something else.

Kripke considers this scenario, and says:

> One cannot repair matters by urging that the subject would eventually respond with the right answer after corroboration by others. First, there are uneducable subjects who will persist in their error even after persistent correction. Second, what is meant by 'correction by others'? If it means rejection by others of 'wrong' answers (answers that do not accord with the rule the speaker means) and the suggestion of the right answer (the answer that does accord), then again the account is circular. (Kripke 1982, p. 32)

We can agree that the notion of correction by others is not going to help matters, for it is entirely possible that someone be persuaded to give up a correct response in favour of a wrong response; and in any case the verdict of the person doing the correcting is also subject to error. But Kripke's first claim - that there are uneducable subjects - is dubious. If I really cannot get someone to add - to recognise that they are making a 'mistake' - then why should we say that they do indeed mean addition? Surely, if someone does not recognise what we should describe as their error as an error, we should accept that they do not mean what we thought they meant. So although we should attempt to analyse correctness in terms of corroboration, Kripke's example which is supposed to illustrate a situation in which the people with the same dispositions mean different things is not compelling. All that is shown is that *these* ideal conditions (the subject is "well rested, in a pleasant environment, free from clutter etc.") are not sufficient to pin down meaning to uniqueness, but that alone does not show that there is no means of doing so. As it is, this example does more to pose the problem (that of specifying ideal conditions) than to demonstrate the futility of searching for a solution.

Boghossian's Holistic Riposte

Perhaps the most promising argument that there can be no ideal conditions is given by Boghossian (1989), which centres on the holistic nature of belief-formation. Boghossian actually sets the problem up using slightly different conditionals from those used here,[22] but

---

[22] Boghossian notes that the ideal conditions are defined as conditions under which the subject is not prone to error - conditions under which any judgement will be correct. Hence, ideal conditions are any conditions O which satisfy the following conditional *a priori* (Boghossian 1989, p. 538):

(BC) For any subject S and concept R: $O \Rightarrow (S \text{ judges } Rx \Rightarrow Rx)$

the underlying principles are the same, namely for a set of O conditions which satisfy the following conditional *a priori*:

(TG) For all S, R: O ⇒ (Jones applies 'R' to x ⇒ 'R' is true of x).

To show that no such set of conditions O exist, Boghossian identifies two additional requirements that O must satisfy in order to be ideal:

> Clearly two conditions must be satisfied [by O]: (i) the specified conditions must really be such as to preclude the possibility of error - otherwise, it will be false that under these conditions 'horse' will get applied only to what it means; (ii) *the conditions must be specified purely naturalistically*, without the use of any semantic or intentional materials - otherwise, the theory will have *assumed* the very properties it was supposed to provide a reconstruction of. (Emphasis added. Boghossian 1989, p. 538)

From this passage it is clear why Boghossian precludes semantic items (claims about what the subject means) from the ideal conditions, for we hope to say what meaning consists in, and so cannot use meaning as a raw material. Yet Boghossian also precludes intentional items in general - that is not only specifically linguistic content, but mental content as well. The reasoning here is that mental content is, from the point of view of the rule-following considerations, on a par with linguistic content. The *concept* of addition, for example, determines an infinite correctness condition just as much as the meaning of the word 'plus' does. (It is correct to believe that 2+2=4, just as much as it is to say it.) If we are in the business of reducing linguistic content to naturalistic properties, we must, presumably, have some objection to the possibility of *sui generis* norms, which means that we will likewise reject the possibility of *sui generis* mental content. For this reason, as Boghossian says, if we specify ideal conditions which refer to intentional mental states, then we seem to have used the very thing which we were supposed to be explaining (i.e. content), and such an account would be circular. That is why the ideal conditions must be specified naturalistically.

Boghossian's conditional concerns the conditions under which the subject is disposed to make correct *judgements*. However, if the aim here is to capture what it is to have a concept R, then it not helpful to refer, within the antecedent, to a situation which requires that S has the concept R: if grasp of the concept R is to consist in a disposition, it cannot consist in the disposition to apply the concept R. Whatever the disposition is that constitutes grasp of the concept, whatever it is that S is *disposed to do*, must be identifiable without reference to the concept R. This does not mean that the disposition must be a disposition to behaviour - it could still be a disposition to judge, so long as the judgement in question is independent of the target concept. Thus the analysis should refer either to the disposition to judge that 'R' is correct, or alternatively to talk in terms of what S *utters*, and what makes his utterances correct, which is the approach used above.

This naturalistic criterion is troublesome, Boghossian claims, because of the fact that we do not make judgements on the basis of experience alone, but also against a web of background beliefs. As Boghossian puts it:

> Neil may come to believe *Lo, a magpie*, as a result of seeing a currawong, because of his further belief that this is just what magpies look like; or because of his belief that the only birds in the immediate vicinity are magpies; or because of his belief that whatever the Pope says goes and his belief that the Pope says that this presented currawong is a magpie. (Boghossian 1989, pp. 539-540)

Generalising, Boghossian notes that "just about any stimulus can cause just about any belief, given a suitably mediating set of background assumptions." (1989, p. 539). If I have beliefs which lead me to disregard the evidence of my own senses, then no matter how 'ideal' my environment is for the appraisal of the question in hand, no matter what sensory information I have, I will not believe the magpie before me is a magpie, and therefore I will not be disposed to call this magpie a 'magpie'. As a result, if it is to be guaranteed that I use my words correctly when the ideal conditions obtain, the optimal conditions must specify that I do not have such interfering background beliefs.

But of course such background beliefs are intentional items. If we can legislate against the potentially infinite number of mediating beliefs within our ideal conditions, then, just because of the requirement that these be specified in naturalistic terms, the dispositional theory becomes redundant. As Boghossian puts it:

> A non-semantically, non-intentionally specified optimality situation is a non-semantically, non-intentionally specified situation in which it is guaranteed that none of this potential infinity of background clusters of belief is present. But how is such a situation to be specified? What is needed is precisely what a dispositional theory was supposed to provide: namely, a set of naturalistically necessary and sufficient conditions for being a belief with a certain content. But of course, if we had *that* we would already have a reductive theory of meaning - we would not need a dispositional theory! Which is to say that, if there is to be any sort of reductive story about meaning at all, it cannot take the form of a dispositional theory. (Boghossian 1989, p. 540)

In sum, because of the holistic nature of verification, we have to include intentional items within the ideal conditions. But a condition on the ideal conditions is that they be specified in naturalistic, non-intentional terms, and so in order to give ideal conditions we need a naturalistic reduction of content. Therefore the analysis of content in terms of dispositions presupposes the very thing which is supposed to be explained, making the dispositional explanation of content superfluous.

In fact, to augment Boghossian's argument somewhat, it can be seen far more easily that the O conditions cannot be stipulated without reference to intentional items. Following a rule, and speaking the truth, are paradigmatically intentional activities. The use of language cannot proceed on a model of stimulus and response, for we do not walk round unintentionally describing everything we see: our utterances are voluntary. Hence, a dispositional analysis must concern the subject's dispositions *given that he intends to speak the truth.* And so it must be stipulated within the O conditions that S has that intention. Intentional items do not have to be included within the O conditions just because of verification holism, but also because beliefs do not produce actions in isolation - the holism of psychological explanation.

An Objection to Boghossian's Argument

Boghossian's argument is the best attempt there is to show for principled reasons that there can be no ideal conditions. Unfortunately it is unsound. This is because there is no need to stipulate that the O conditions be non-intentional. Certainly, the example which Boghossian uses to prove his point does not show that it is circular to include intentional items within the O conditions. In the example, the concept which is analysed in dispositional terms is the concept *magpie*. However, the concepts mentioned in the O conditions concern the Pope and currawongs, not the concept magpie. Thus, in referring to propositional attitudes involving these concepts we have *not* assumed the property that is analysed with the conditional. What we *have* assumed is that Jones has the property of believing what the Pope says is true, etc.; what we have *constructed* is the property of believing that x is a magpie. These are different properties. In giving an analysis of the property magpie we refer to the concepts currawong and Pope, but without comment as to their constitutive natures. So, including such items within the O conditions does not stipulate, or entail, that grasp of the concept magpie is dispositional. In the ideal conditions we do have to refer to intentional items, but if these are not the *same* items as those that the conditional is supposed to capture, then there is no circularity.

Nevertheless, in referring to any concept whatsoever within the O conditions have we not, as Boghossian claims, assumed a dispositional account of content? Frankly, no. It is quite legitimate to refer to certain intentional items as required within the O conditions without regard to their constitutive nature, and hence without *assuming* that a dispositional account of content is possible. In referring to a concept we do not assume that grasp of the concept is dispositional. The project is to give a dispositional analysis of some new intentional item. In

order for this analysis to be non-circular, all that is required is that possession of the new intentional item be independent of possession of the concepts mentioned in the O conditions. The independence condition ensures that the new item is constituted by the disposition in question, and is not tacitly assumed within the O conditions. If the analysis of the new concept is successful - that is if at least one concept is dispositional - then there is no reason to think that the concepts referred to within the O conditions should not also admit of a dispositional analysis. In that case it is not *assumed* that the concepts referred to within the O conditions are naturalistically respectable - it is *discovered* that they are. The problem only arises on the prior assumption that there must be a *naturalistic* account of content, which is manifestly not the case.[23]

Although Boghossian's argument fails as given, there remains a certain amount of scope for development. For example, we might note that in order to apply the word 'magpie' to a particular bird, we need have certain background beliefs about currawongs; and likewise, to give a dispositional analysis of the meaning of 'currawong', we should have to preclude certain background beliefs involving the concept *magpie*. As a result, we could only give an analysis of the concept *magpie* by referring (*via* the concept *currawong*) to that very concept within our ideal conditions. So perhaps the account is circular after all.

What is notable about this situation just described is that we here consider - in contrast with the examples used by Kripke - high-level, theoretical terms. It is not at all obvious that with a restricted set of beliefs, about more basic properties (such as colour, shape, and the like) that the same type of situation would arise. In fact, we cannot be certain that such holistic considerations arise at the higher level without a better understanding of the conceptual hierarchy involved.

The overall message is again that there may well be difficulties that the dispositional thesis has not yet overcome, but there is nothing to say that they cannot be overcome, and so have not established that ideal conditions cannot be given. The second strand of Kripke's argument against dispositions does not achieve its aim either.

---

[23] It is noteworthy that whereas Boghossian treats dispositionalism as a naturalistic thesis, his own characterisation of the thesis is given in terms of the disposition to *judge*, so that it automatically refers to intentional items. Although dispositional accounts of content are familiar in the context of dispositions to behaviour, there is nothing amiss, even from a Cartesian standpoint, in talk of dispositions to judge, or to feel pain, or whatever, where judgements and feelings do not admit of a naturalistic analysis. Dispositional theories are not naturalistic *per se*.

The Irreducibility Thesis

Leaving aside dispositions for the moment, the second contestable part of Kripke's overall argument I wish to concentrate on is his dismissal of meaning as a primitive, irreducible state, and of Platonism. As we have seen, in both cases the requirement is that a 'finite' mental state determines an infinite extension (in the one case directly, in the other via a Platonic object). Kripke objects that any 'finite' mental state can be variously interpreted, and so fails to determine a unique extension.

To assess this type of argument, it is important to recognise the distinction between the fact that something *can* be variously interpreted, and the fact that it *needs* to be interpreted to fix an extension. It can be accepted without difficulty that any object whatsoever *can* be interpreted in many ways. If interpretation means correlating an object with a correctness condition, then we can do this in an entirely arbitrary manner. However, to say that an object can be interpreted is not to say that it stands *in need* of interpretation. Specifically, the suggestion that meaning is a *sui generis* state is motivated by the thought that such a state may be intrinsically normative, that is that by its very nature it sets a standard for correct behaviour, and so does so without recourse to interpretation of any kind. In that case, the fact that it itself determines one unique correctness condition is not undermined by the claim that it *could* determine other correctness conditions under various interpretations. The correct extension would be the one extension fixed intrinsically.

This rebuttal clearly rests on the claim that a 'finite' state of mind cannot determine an infinite extension, and perhaps the thought underlying Kripke's position is that to get from the finite to the infinite, interpretation is essential - so the regress does arise after all. However, the ability of a finite state to determine that each one of an infinite collection of answers is correct - and to do so without interpretation - should not strike us as *especially* troubling. Suppose I show you a picture of a (generic) tree, and ask you to go and find one. There is no logical reason why there should not be an infinite number of trees, with the picture being, in a perfectly ordinary sense, a representation of each one of them. Now the actual picture (the marks on the page) seems to be finite (at least it exists on a finite piece of paper), and yet it represents an infinite number of objects.

A physical object such as a picture of a tree is not intrinsically representational, for it depends on the mind to bestow it with representational power. However, this fact does not

make the example any less relevant as a countermeasure to Kripke's observations, for it is precisely the ability of the mind to represent things which answers Kripke's concerns about infiniteness of meaning. Indeed, the power of representation is the essence of any intentional state. It is especially noteworthy that intentional states such as intentions do not stand in need of interpretation in order to represent, they are intrinsically representational. Correspondingly we ought not entertain for a moment the thought that the mental representation necessary for meaning leads to a regress of interpretations.

Given that intentional states in general are intrinsically representative, and representative states can represent an infinite number of objects, then we have no reason to find the infiniteness of meanings particularly vexatious. A normative meaning state must at least be representational; it must be 'about' its correctness conditions in the way that an intention is 'about' its satisfaction conditions. This is not to say that mere representation is the same as normativity (a picture of a tree does not determine a correctness condition, for example). For normativity we require further that the actions represented by the rule must be represented *as correct*. However, it is not the ability of *sui generis* states to determine correctness (to represent actions as correct) which is at issue, but the ability of a state within a finite mind to determine an infinite correctness condition. Once representation is secured, there is no reason to think we cannot have representation as correct.

It remains true, of course, that the ability of a state to represent without interpretation is somewhat mysterious; but mysteriousness does not warrant the conclusion that such states do not exist. In fact, far from producing any reasons to deny the existence of such states, Kripke's argument is blocked by the possibility of them. As previously suggested, we should expect any conclusion from the 'sceptical' argument concerning our ability to put conceptual content into our utterances to apply in a corresponding manner to concept possession in general, even in the absence of linguistic expression. Both linguistic content and mental content require grasp of the same correctness conditions, and so any argument focused on the notion of a correctness condition is likely to apply in both cases. Yet the only thought Kripke produces against the possibility of *sui generis* concept possession is that it is mysterious. In this way Kripke mis-locates the onus of proof. We ought not think that grasp of a rule (or meaning) is problematic just because it requires us to pack the representation of the infinite steps determined by the rule into a finite mind. Instead, we should see that the infiniteness of rules is *unproblematic* precisely because of the infinite character of finite representations. A single representation may represent numerous, indeed an infinite number, of distinct objects,

and so our ordinary notion of representation allows us to overcome the infiniteness problem.[24]

The same type of observation may be used to counter Kripke's rejection of the Platonic theory of rule-following, but there are additional weaknesses in this part of his argument. As we have seen, whilst Kripke accepts that in such an account the Platonic object can determine a correctness condition without interpretation (1982, p.53), he does find the manner in which the mind latches on to such a state to be troublesome. He says:

> The idea in my mind is a finite object: can it not be interpreted as determining a quus function, rather than a plus function? (Kripke 1982, p. 54)

Again, the question arises: is it that such a state *can* interpreted, or is it that it *must* be interpreted? To make the argument persuasive, it has to be established that interpretation is essential.

Presumably the Platonic object in question can be accepted as being infinite - in that it determines an infinite extension intrinsically - whereas the mental state is finite, so again it is in bridging the gap between the finite and the infinite that the need for interpretation plausibly arises. Yet it has already been shown that a finite state can determine an infinite extension without interpretation, and so this type of consideration could be put to good effect here also. In that case, interpretation would not be an essential part of the account, and the regress need not arise.

In addition, the difficulty Kripke finds here is dissolved when we enquire into the way in which the mind grasps an abstract object. It is difficult to know what we are to make of this 'grasping' function, but the obvious analogy is with the way in which a hand takes hold of a physical object. If the analogy is remotely apt, then since a hand may quite easily grasp an infinitely long rope simply by grasping any one part of it, then we should likewise accept that the mind may readily grasp an infinite Platonic object. The point is that the notion of

---

[24] Perhaps one reason why the infiniteness condition appears more formidable than it really is rests with Kripke's choice - or rather Wittgenstein's choice - of a mathematical example to illustrate the problem. In such a case the idea that a representation of a function has to be an infinite table is reasonably attractive (as opposed to thinking for example that grasp of the concept green requires a list of all green things). There is no reason why the suggestion that an infinite list must be stored in the mind should be seriously entertained. We actually carry out large computations by following an algorithm, and so a more plausible proposition is that we internalise a representation (an intrinsic representation) of the algorithm, and that an infinite number of calculations - those produced in accordance with the algorithm - are represented by this mental state.

grasping is precisely the type of relationship which could enable the gap between finite and infinite to be bridged.[25]

Conclusion

The result of the preceding discussion is that Kripke's elimination of putative meaning-determinants has failed to exclude some important contenders, namely dispositions, *sui generis* states, and Platonic objects. *Contra* Kripke, it has not been shown that dispositions are inadequate when it comes to constructing correctness, nor that sui generis state and Platonic objects succumb to a regress of interpretations. All three options are, in fact, very much live positions, and the fact that all three remain is a strong indictment of the 'sceptical' argument.

---

[25] The *sui generis* response to Kripke's sceptic is developed by Wright (who develops the analogy with intentions) (1984 and 1989b), is mentioned by McGinn (1984), and is the position Boghossian advocates as the conclusion of his (1989).

2. Following a Rule

In the previous chapter we noted that, although termed 'sceptical', Kripke's attack on rule-following actually engages with distinctly ontological matters. Since that approach fails to raise a substantive threat to the possibility of rule-following, I want in this chapter to consider whether the 'missing' epistemological dimension is any more fruitful. A true 'sceptical' argument would aim to show (a) that there is some body of knowledge we think we possess, which in fact we do not, and (b) that this lack of knowledge proves damaging to our view of ourselves as rule-followers. Is there any such argument?

To lay things out fully, there are three types of knowledge which I possess as a rule-follower, and which are relevant to the discussion. These are:

Existence: I know that I am following a rule.
Identity: I know that I am following rule R.
Application: I know that the rule I am following now requires φ.

Questions concerning the existence, and the identity, of the rules we follow have both been raised in connection with Wittgenstein. For example, Kripke briefly asks how we know the identity of the rules we follow, and as we shall see in Chapter 5, Wright has developed a forceful thesis on the basis that such knowledge is not as straightforward as we might think, and indeed uses such epistemological issues to motivate a substantial revision to the notion of rule-following.[1] For a complete account of the epistemology of rule-following, this kind

---

[1] In the previous chapter I claimed that Kripke's 'sceptical' argument centered on ontological, not epistemological, issues. Somewhat exceptionally though, Kripke does raise this objection to the *sui generis* state theory:

> [The *sui generis* meaning state] is not supposed to be an introspectable state of affairs, yet we supposedly are aware of it with some fair degree of certainty whenever it occurs. For how else can each of us be confident that he does, at present, mean addition by 'plus'? (Kripke 1982, p. 51)

As Kripke notes, any meaning-determining state cannot be essentially phenomenological, for a phenomenological state is not normative. But if the meaning-determinant is not phenomenological, how do we know what we mean by our words?

At this level of development it is not possible to see this objection as particularly powerful. This is because here Kripke raises an issue which applies to all propositional attitudes, not just meaning. In a great many cases we

of argument has to be examined, but for structural reasons I have deferred scrutiny of this type of consideration until later. This is because it is not clear that knowing either that I am following a rule, or that I am following a particular rule, is necessary for me to be a rule-follower, which is to say a threat to either type of knowledge is not a threat to rule-following itself. From this initial appraisal, it seems that a demonstration that such knowledge is unavailable would not have the type of destructive consequences which currently hold our interest.

The situation is quite different when it comes to matters of application. The very term 'rule-following' indicates that the rule is something which is *followed*. In following a rule I *read off* the requirements of the rule; the rule *informs* my actions; it *guides* me. So, on the face of it at least, a rule must be the type of thing with which one can have epistemic contact. And if it can be shown that we cannot attain the necessary epistemic contact – if we cannot know what the rule requires of me in any given situation – then we cannot be rule followers. A failure of knowledge with respect to the application of a rule would then appear to lead to the same 'paradoxical' position as was claimed on the basis of the Kripke's 'sceptical' argument.

Kripke's failure to prosecute the epistemological dimension of rule-following is the most significant lacuna in his account of Wittgenstein's argument, for the issue was of central importance to Wittgenstein. To assess whether there is any merit in this type of approach, I want to initially look at the culmination of Wittgenstein's epistemological argument, which occurs in the following passages of the *Philosophical Investigations*:

> 217 "How am I able to obey a rule?" - if this is not a question about causes, then it is a question about the justification for my following the rule in the way I do.
> If I have exhausted the justifications I have reached bedrock, and my spade is turned. Then I am inclined to say: "This is simply what I do."....
>
> 218. Whence comes the idea that the beginning of a series is a visible section of rails invisibly laid to infinity? Well, we might imagine rails instead of a rule. And infinitely long rails correspond to the unlimited application of the rule.

---

know with complete certainty what our beliefs and desires are, and yet it is far from obvious how this knowledge is obtained. Notably, though, we do not find the fact that beliefs, intentions, and the like are introspectable, but non-phenomenological, sufficient grounds on which to deny that such states exist. Correspondingly, the fact that a *sui generis* meaning would also have this property is not sufficient reason to say that there can be no such state. What holds for intentions should hold for meanings: although we can produce no convincing account of how we come to know what we mean, our ignorance in this respect ought not lead us to deny that meaning is a *sui generis* meaning state. In short, in arguing against a *sui generis* meaning state, Kripke does nothing to identify a particular problem with rules, meaning, or correctness, but rather one of the characteristics of such state. In contrast, Wright does offer a sustained and far more compelling discussion of the connection between knowledge of meaning, introspection, and the way in which such factors may have ontological significance.

> 219. "All the steps are already taken" means: I no longer have any choice. The rule, once stamped with a particular meaning, traces the lines along which it is to be followed throughout the whole of space. - But if something of this sort really were the case, how would it help?
> No; my description only made sense if it was to be understood symbolically. - I should have said: *This is how it strikes me.*
> When I obey a rule I do not choose.
> I obey the rule *blindly*.
>
> 220. But what is the purpose of the symbolical expression? It was supposed to bring into prominence a difference between being causally determined and being logically determined.
>
> 221. My symbolical expression was really a mythological description of the use of a rule.

At the core here there is a blueprint for the type of argument we are after: if we are to think of rule-following as involving cognitive contact with a rule, then the rule-follower must have a reason for acting as he does. And if no such reason is available - if all reasons really are 'exhausted' - then the model involving cognitive contact falls down.

With a view to assessing the destructive force of this argument, it is notable that (at least in the quoted passage) Wittgenstein does not present this as an argument against rule-following *per se*. (The conclusion is not that rule-following is impossible, only that when following a rule we act 'blindly'.) In light of this it seems reasonable to take the argument as merely an advocation for a revision to the epistemology of rule-following: whereas we think we are guided (by the rule) when we follow a rule, in fact we are not.[2] I myself doubt that Wittgenstein intended the result to have such limited scope, but certainly if the eventual message is to be more substantial, additional argumentation is required - namely, to show that such an epistemological revision is itself unstable. And no matter quite what Wittgenstein's views on the matter were, the important point is that this instability does indeed occur, meaning that the 'exhaustion of reasons' argument *is* destructive of rule-following. To see why this is, and to appreciate the significance of the argument against the existence of reasons fully, it is necessary to look at the overall structure of the argument given in the passages quoted above in more detail.[3]

---

[2] Malcolm (1986), for example, takes the view that the message of these passages is nothing stronger than the counter-intuitive result that rule-following does not involve being guided.

[3] One point which has some bearing on the eventual message of the passage is the significance we accord to the picture of the 'rule-as-rail'. This has been taken to be a reference to Platonism, with the section as a whole being directed against a Platonic philosophy of language. On this reading the 'rail' is identified with an object in the Platonic realm which determines the correct use of a word. I find this reading to be strained, for the analogy is only brought in after the central conclusion - that we act without reason - has been reached. Indeed, the advertised target of the argument is the very existence of a guiding infinite correctness condition (PI §219), a view which is not explicitly Platonic. Certainly Platonism is one of the intended targets - a Platonic rule only enters the picture

Wittgenstein's stated aim is to explain how rule-following is possible, which means to explain how we systematically (if not infallibly) succeed in according with a rule. Whilst this is ultimately an account of the way we *act*, to explain these actions we need do no more than explain how we make the right *choice* of response. Given that following a rule is an intentional act, my φ-ing is explained by (a) my intention to follow rule R, (b) my belief that R requires φ in C, and (c) my belief that condition C now obtains. The rule enters the explanation in so far as it explains (b), that is how I come to believe that the rule now requires φ. So to explain how it is possible to follow a rule, the search is for an explanation for my *choice* of action - why I believe that a given action is, in present circumstances, correct, and how my choices tend to be correct.[4]

Initially, we assume that the norm (or normative mental state), and one's opinion as to the requirements of the rule, are independent. That is just to acknowledge that the rule determines what is correct in advance of any thoughts I have on the matter, and as a rule-follower my job is to accord with this pre-established standard. The aim is to explain how the rule influences the verdicts I give. To this end, Wittgenstein considers three possibilities, marked by three different relationships between rule and verdict: a causal explanation, a reason explanation, and the possibility that there is no relevant relation at all ('blind' rule-following).

Only the first two options are at all explanatory, and the role of the rule in each explanation is significantly different. If we are talking in terms of causes, then we can only mean that the subject's apprehension of her worldly situation C causes her to believe that φ is correct. If a rule is to enter into this account, then it must be that the rule is, not a cause itself (for a rule is

---

in as much as it guides one's actions. But given the doubts expressed above about the very possibility of 'blind rule-following', the force of the point is not necessarily given such a sharp focus.

[4] This identification of what is to be explained is important not least because Wittgenstein himself vacillates between explanations of (variously) behaviour, action, and belief. For example, in PI §211 he starts by asking how someone knows what is required (reason for *belief*), and ends on an observation about what the agent then does (no reason for *action*). Whilst the explanations for such phenomena may well be closely related, it is also possible that types of explanation may be available for one category and not for another. I have in mind the way Wittgenstein discounts the possibility of *causal* explanations for certain rule-governed phenomena (see for example PI §169), and certainly the treatment we give to causal explanations for beliefs, actions and behaviour may be significantly different.

not an event), but rather a causally relevant factor - it is only because the subject grasps a rule that her apprehension of C causes the relevant belief.[5]

On the other hand, if the explanation comes in terms of reasons, we should expect both worldly situation and the rule each to be a part of the overall reason for the given verdict. Here it is because situation C is appraised, and because R requires φ in C, that the subject believes that φ is correct.

Wittgenstein rejects both causal and reason explanations, leaving no explanation for rule-governed behaviour. Instead, we merely act in whatever way strikes us as correct. Although the expression used to describe this result - 'blind rule-following' - has the air of an oxymoron (to say you act blindly is to say that you are not following anything at all), the notion of 'blind rule-following' cannot be dismissed automatically. We initially think that rule-following involves some kind of epistemic contact or guidance, which makes an epistemic aspect at least desirable, but it is not obvious that this element is absolutely essential. In particular, a dispositional theory of rule-following would hold that a rule-follower does what seems right without justification. Indeed, if in following a rule I act without reason, and without a casual relationship with a rule/grasp of a rule, then following a rule must consist in the disposition to form certain verdicts, and correctness is (presumably) to be constructed out of these responses. That is, 'blind' rule-following collapses into a form of dispositionalism.

It is for this reason that the passage quoted has been taken as an indication that Wittgenstein endorsed a dispositional theory. This interpretation is not tenable, however, for Wittgenstein explicitly argues against dispositionalism (PI § 149). The prudent course is to see PI §§ 217-221 as part of a larger argument: this stage establishes that all that rule-following can be is dispositional, with other considerations being raised elsewhere to refute this thesis.

We have already rejected the argument against dispositionalism which Kripke derives from Wittgenstein, and in terms of exegesis, Kripke's interpretation is not here deficient, for there are no further elements presented by Wittgenstein which would strengthen the case presented by Kripke. However, once we fully engage with the epistemic dimension of rule-following,

---

[5] It should be noted that the claim here is only that the rule has a causal role; it is not the functionalist claim that grasp of a rule *consists* in a causal role.

there is, I think, a means of refuting dispositionalism altogether. In fact, the argument is not specifically against a dispositional thesis, but is also effective against any position which dispenses with reasons (so it is effective against the theory which utilises a causal explanation as well). Before giving the argument, I want to show that the attendant refutation of the causal account is in itself significant, for although Wittgenstein rejects such a theory, he does not properly motivate such a course of action.

The outcome will be this: of our three accounts of rule-following, the argument to be given will refute both the causal and dispositional ('blind') theses. As a result, Wittgenstein's conclusion that we follow a rule without reasons becomes untenable - neither option which dispenses with reasons is now available. This means that the core argument that we act without reasons becomes an argument that rule-following is impossible. To get to this stage, we need first look at the causal explanation, and then at the case against both 'no reason' theories.

Causal Explanation

In PI § 217 (quoted above) Wittgenstein merely puts the possibility of a causal explanation of rule-following to one side, for it is an options already rejected in preceding sections of the *Investigations*. Taking our cues from PI §§ 169, 195, and 198, the reason for rejecting this thesis is that:

(a) my worldly situation is a reason for my following the rule in the way I do

(b) a reason is not a cause.

If my worldly situation is a reason for my belief that φ is correct, and not a cause, then the causal explanation of rule-following is not available.

Certainly, we may accept that the situation I perceive myself to be in is a reason for my following the rule in the way I do. My situation is within my awareness, and would be cited by me as a relevant factor in my decision to act as I do. What is far more contentious is Wittgenstein's claim that a reason is not a cause. The thesis is forwarded on the basis that the receptive epistemologies are quite distinct - it is *a priori* that we know what our reasons for our actions/beliefs are, but we can usually only establish a causal connection by way of empirical observation.

There are two reasons to treat this premise in the argument with caution. One is Davidson's (1980) argument that a reason *must* be a cause if the notion of a reason is to carry any explanatory value. I do not propose to explore this particular debate, but will merely record that the issue is the subject of on-going contention.

The second difficulty with Wittgenstein's argument is that even if we accept that a reason cannot be a cause, it may still be possible to combine both reason-explanations and causal explanations within a single explanatory framework for a given phenomenon. For example, suppose that my seeing a red book is my reason for believing that φ is correct. It might be that seeing the book caused me to accept that the presence of the book is a reason to believe that φ is correct. In other words, it might be that the same situation is both the reason for one thing (my belief), and the cause of something else (my acceptance of a certain reason for belief), and that both elements contribute to the overall explanation of why I formed the belief that I did. In this way we can accept that a reason explanation is not a causal explanation, but still incorporate both causes and reasons in one explanation.

An investigation into whether this type of observation is relevant to Wittgenstein's dismissal of causal explanations is not a project worth perusing here. For the moral is merely that Wittgenstein's rejection is insecure, and so the argument which follows which does preclude such a thesis about rule-following is performing a genuine task.

The Need for Reasons

To bring out the difficulty faced by any theory of rule-following that does away with reasons I shall look at the dispositional theory, and then show that the problem it faces applies more generally. One of the chief merits of the dispositional theory is that it not only states what grasping a rule consists in, but it also automatically accounts for the way that the requirements of the rule come effortlessly to mind. Following a rule consists in doing what strikes one as correct, so there is no substantive issue arising as to how one arrives at one's verdicts.

Unfortunately, these two elements - that *knowing* the requirements of a rule is a dispositional matter, and *following* a rule is a matter of doing what seems right - are incompatible. The conflict emerges when we consider a situation in which a rule-follower - call him D - comes to believe in the dispositional thesis itself. By hypothesis, the theory is a correct description of the world, and so ought not unduly interfere with the subject's otherwise correct

description of his worldly situation. Yet adoption of the theory may actually have a catastrophic effect on D's classificatory practices.

As a rule-follower, D must of course aim at correctness - to try to do what the rule requires in any given situation. Suppose, though, that upon familiarisation with Wittgenstein's 'exhaustion of reasons' argument, D comes to acknowledge that he actually has no reason to believe that any particular verdict is correct, and that all he does when following a rule is manifest certain dispositions. As we have seen (Chapter 1), the dispositional thesis requires that correctness be constructed out of what seems right under ideal conditions. Given that D believes the dispositional thesis, and so knows how dispositions are supposed to yield truth, the obvious means for D to aim at correctness would be for him to (a) identify the correctness-determining ideal conditions, and (b) ascertain what he himself would do were he in those ideal conditions. By aiming to do what he would ideally do, D would thereby be aiming at what is correct. Certainly this procedure is beset with difficulties: as we have seen, the identity of the ideal conditions is not something which can be readily ascertained; in addition, it may be difficult to discern when such conditions are met; let alone to discover what you would do were you (counterfactually) under such conditions. Yet the fact remains that this is the strategy which anyone aiming at correctness would be justified in pursuing.

In fact, given that the result of any such enquiry may not be immediate, in the meantime D would be rationally warranted in withholding all verdicts as to what his rule requires. Of course, aiming at correctness does not require omnipotence, or infallible powers of reasoning, so the occasional failure to acknowledge reasons for or against a belief, or making an erroneous inference as to the significance of some 'evidence', does not undermine one's status as a rule-follower. Yet if D comes to acknowledge a theory of rule-following under which it is accepted that he *systematically* fails to have any reason whatsoever to believe that φ is correct, then a rational response would be a global agnosticism: to withhold any judgement on the question in hand until more information is in.

Before identifying the difficulty inherent with this situation, note that the causal theory entails a similar result. The theory this time is that sensory inputs cause certain beliefs as to what is correct, and that the causal connection between input and belief is mediated by a rule. Given that this type of account must accommodate the possibility of error, we should accept mistaken verdicts are the result of some kind of disruption to the 'normal' causal process. In that case, again, there will be a distinction between ideal and non-ideal causal

conditions, and consequently the obvious way to aim at truth would be to identify those conditions, and to do what you would do were those conditions satisfied.

Whether the investigation into ideal conditions is instigated in the name of either the dispositional or the normative/causative theory, anyone adopting this strategy would certainly be someone trying to accord with the requirements of a rule, but they would *not* count as a rule-follower, for they could no longer be described as someone who *knows* the requirements of the rule. The essence of rule-following is that the rule's requirements are internalised, and subsequently govern one's behaviour. In contrast, our example of someone theorising about how the rule is to be applied is manifestly someone who does not have knowledge of the rule, but rather is someone intent on *discovering* what its requirements are. So, although aiming at correctness in some sense, D is not then someone who uses his dispositional knowledge as a (dispositional) rule-follower should.

In response, the dispositionalist has little option but to claim that although investigating the constitutive nature of truth is *one* way of aiming at truth, the point of the dispositional theory is to accept that simply doing what seems right *also* counts as aiming at correctness - and further, that this sense of aiming at correctness is all that is required for rule-following.

However, such a view is not tenable, for this latter attempt to capture the notion of aiming at correctness is inadequate. To see this, we may ask in general terms, what "aiming at F-ness" means. Certainly a minimum requirement is that the subject intends to do what is F. In addition, they should take into account any relevant indicators which suggest that an action is F, or indeed that it is not F. But in addition, we should expect an appropriate reaction to the *absence* of any such indicators: *if you know you have no reason to believe that φ is F, then do not accept that φ is F.* For if you know that your belief is unjustified, if you have no reason to believe that φ is F, then you cannot rationally forward that opinion as an attempt at F-ness.

There are perhaps situations in which an absence of justification for an action does not undermine that action. For example, suppose that someone who is trying to throw balls of paper into a waste-paper basket is reliably informed that he has been subject to an illusion, and that his visual inputs give no good indication of where the basket is within the room. If his aim is to get a ball in the basket because he wins a prize, then he may as well throw as he

would have done otherwise - any action gives him a better chance of winning than no action at all. The 'blind' action is valid in the hope of success as a matter of chance.

But aiming at correctness is not like this example, in that the urge to act based on the chance of success ought not override the urge to withhold action given the possibility of failure. Rather, when following a rule, the need to avoid failure (i.e. to act incorrectly) is *as important* as the need for success. Haphazard truths are not what we are after; the predominance of truth over falsity is. In this type of situation, we are not just hoping for success, but actively trying to avoid error. Failure is undesirable, not simply because it marks an absence of success, but as an intrinsic characteristic of the given endeavour. And it is precisely this type of situation in which we have to take the lack of indicators - the lack of reasons - seriously. Pot luck is no good, and so anyone aiming at correctness cannot carry on regardless of their lack of justification.

We cannot intend to speak the truth and acknowledge we have no reason for our beliefs as to what is true: to do so is simply an admission that you are not aiming at all. What we have discovered is that in certain situations both conditions - aiming at truth, and doing what seems right - are mutually exclusive. It is because of this incompatibility that all 'no reason' theories are to be rejected.[6]

Exhausting Justifications

In demonstrating that a rule-follower cannot act without reason, we have dismissed both of our flanking positions - a causal theory and a dispositional ('blind') theory. Our focus now falls squarely on Wittgenstein's central contention, namely that in following a rule all reasons are 'exhausted', and we are left to act without reasons. If this is right, then given our rejection of the other options, rule-following is indeed impossible.

The key issue here is in what sense justifications for one's choice of action are 'exhausted'. One interpretation would be a quite general claim about the nature of justifications, that justifications come to an end because eventually it is not possible to justify the claim that a justification is a justification. For example, suppose I say that Bill has measles, and make my

---

[6] It might be countered that belief in a 'no reason' theory be ruled out by the ideal conditions of the dispositional theory - belief in the theory being exactly the type of background belief which interferes with the production of correct actions, and which Boghossian notes has to be excluded if the theory is to be tenable. However the reason for excluding background theories was that false beliefs can produce false verdicts. The adoption of an *ex hypothesi* true theory cannot be discounted in this way.

claim on the basis that he is covered in red spots. In this particular situation it is a fair question to ask why Bill's being covered in red spots is a justification for the claim that he has measles. And if I can provide no answer, then a reasonable view is that I am not justified in claiming that he has measles on the basis of his spots. That is, *in some circumstances*, a justification must itself be justified as a justification in order to be justificatory. But clearly, if this were *always* the case, then the notion of justification is regressive: the theory in question holds that a justification is only justificatory given a second justification; but then that second justification will only be justificatory given a third one, and so on. If this were a proper account of justification, then at some point we would always reach a justification which could not itself be justified, a situation which would leave us without justification, and which could be described as a situation in which our justifications are "exhausted".

This regress can only be stopped (and in order for anything to be a justification it must be stopped) if some justifications do not need further justification in order to be justificatory. It is just this type of argument which motivates foundationalist epistemologies, which state that certain statements which can be known (and are thus justified) without recourse to further justification.

Taking our example above, we might cite as evidence for our claim the fact that every previous patient manifesting such spots has been found to have the measles virus, and so the justification works by induction. Whether we want to say that induction stands as a type of justification, or if we choose to justify induction using deduction, the end point will, presumably, be *a priori*, and need no further justification. Similarly, the justification we might offer for an observation statement might eventually be given in terms of the contents of our experience, and no justification is required for the claim, say, that there is a red spot in my field of vision.

According to Baker and Hacker (1985), it is precisely this limitation to the extent in which any justification can itself be justified that Wittgenstein is concerned with. The conclusion as applied to the application of a rule is the same as it must be for justification to be possible at all: there must be some stage at which a supporting justification is not required:

> Wittgenstein's point is not that here my action is unjustified (haphazard, a free choice), but rather that it has already been justified, and no further justification stands behind the justification which has been given. (Baker and Hacker 1985, p. 209)

The "justification which has been given" that Baker and Hacker refer to consists in how I have been trained to follow a rule, and the sample given. The claim is that it is disingenuous to search for any further justification, and we are only tempted to do so because we fail to recognise that our training is justification enough.

This interpretation of the argument is unsatisfactory, since the whole motivation behind the rule-following considerations is that past performance cannot justify the choice of one action as being in accordance with the rule over any other. Past performance could justify *any* current choice of action. Moreover, this reading is not true to Wittgenstein. He does not conclude that the justification I do have is sufficient justification; what he does say is that when I follow a rule "I act, *without* reasons" (Emphasis added, Wittgenstein PI §211). To say that I act *without* justification is quite different from saying that the justification on offer is adequate. The end of justification is not presented as the result of a bogus requirement for a further level of justification. Rather, in the case of rules, there is some judgement which we think of as justified, but which genuinely fails to be so. For this reason the correct account of the argument cannot be that the "exhaustion" of justifications is due to any such general regress.

The Regress Argument

An alternative, and ultimately more satisfactory, reading of this passage equates the 'exhaustion of justifications' with the 'regress of interpretations' emphasised by Kripke. Such a reading is given by Pears:

> The weight of Wittgenstein's attack falls on the...suggestion that there is something that actually occurs in his mind and gives him infallible guidance. This thing, whatever it is, is supposed to lock him onto the fixed rails of correct use because its meaning is instantly self-intimating. It is the kind of thing that qualifies as an instant mental talisman.
>
> But what can it possibly be? As we have seen, mere words will not do the trick because any analytical formula will itself stand in just as much need of interpretation. Is it, then, a mental image or picture? But though a picture may have a natural application, it will also have many other applications, any one of which could be chosen instead of the natural one. It soon becomes apparent that Wittgenstein has a general objection to any candidate for the post of instant mental talisman: nothing could fill the post, because any single thing in anyone's mind could always be connectable with more than one set of applications. (Pears 1988, p. 469)

And he sums up:

> All that could possibly be found [to be a mental talisman] is...an equivocal image or formula that itself needs to be interpreted. (Pears 1988, pp. 469-470)

The argument here is reminiscent of Kripke's 'sceptical' argument in that it works by exhaustive elimination and vicious regress. The significant difference is that whereas Kripke claims that no phenomenological episode can determine a norm, here the thought is that no such entity can provide a suitable *justification* for a judgement as to what is correct. Nevertheless, it turns out that the items which featured on Kripke's list fail *as justifications* for much the same reasons they failed as rule-determinants. It is by now a familiar point that any finite number of actions fall under an infinite number of rules. But then, if prior use is all we have to appeal to, any action which I decide accords with the rule in the present is justifiable on the basis that it continues 'the same' rule as before. Therefore prior use is insufficient as a justification for present action. And if a qualitative mental state is to indicate what action is correct, then this can only be on the basis of an interpretation. Just as it is not an intrinsic property of, say, a mental image of a square to determine a correctness condition, it is not an intrinsic property of such an image square that it *tells* me to apply the word "square" to squares. What I need is an interpretation, a further rule such as: when an image of a square occurs, say the word "square". But then the justification provided by the image rests on my ability to follow a rule (the interpretation). If I need a justification for following a rule in the way I do, then clearly I need a justification for following the interpretation in the way I do, and this leads to a regress of justifications, and a regress of interpretations. The regress of interpretations therefore applies just as much at the epistemic level as it does at the constitutive level. Yet, if no behavioural episode, and no phenomenological episode can justify my choice of action, all the options appear to have been exhausted: we have run out of justifications.

Self-Justification

In assessing this argument, we may ask: are the items considered (experiences, instructions, formulae) the only justifications on offer? The argument proceeds by a process of elimination, and so the list of putative justifications ought to be all encompassing if the reasoning is to be sound. Yet there is a notable omission, which emerges when we consider the way in which statements about many mental states are usually considered to be justified. As Shoemaker says:

> It is the distinguishing characteristic of first-person experience statements...that it is simply their being true, and not the observation that they are true, or the possession of evidence that they are true, that entitles one to assert them. (Shoemaker 1963, p. 122)

Whilst we should often think that the justification for a given statement is distinct from that which makes it true, in the case of qualitative mental states (pains, itches, and so on) these

are identical: the same state of affairs both justifies a statement and makes that statement true. Indeed, this is not only plausible in the case of qualitative mental states, but holds for propositional attitudes as well. We ordinarily think that we know the contents of our own minds directly, by introspection,[7] and that statements about our contentful mental states do not stand in need of any justification: I know what I believe in virtue of my believing it. It is thus a characteristic of statements about the mental that they may be self-justifying, by which I mean that the following conditional is true:

$(\forall S)$ S is F $\Leftrightarrow$ S is justified in believing she is F

Can the same be said of the judgement that φ accords with my rule? As we have seen, the suggestion which defeated the 'sceptical' argument was that grasp of a rule is a *sui generis* mental state. If grasping a rule is a mental state, then it is reasonable to suppose that we do indeed know what our rules require in the way that we know what we believe and intend. In that case, the claim that I should now φ is justified by *the fact* that φ accords with the rule I am following. If so, then certainly I do not act "without reasons", nor "blindly", in the same way that I do not identify my sensations "blindly". I say that I have a pain simply because I do have a pain: by analogy the reason behind my decision that the rule requires φ is, simply, that the rule requires φ. In that case, the conclusion that we act without justification can only be reached if we persist in our requirement that the justification for our action be *distinct* from the truth-condition, a requirement which is wholly without foundation. To search for any other justification, or any further explanation would be to betray a misunderstanding of the nature of the case: the fact justifies the claim, and that is why no justification in any other terms can be given. If this is right, then Wittgenstein is simply guilty of overlooking the most obvious justification, the fact itself.

Self-Justification and the Possibility of Error

The idea that rule-following involves self-justifying beliefs faces one difficulty, in that the account appears to make inadequate allowance for the possibility of error. We should first

---

[7] The passage quoted from Shoemaker actually occurs in the context of an argument directed against the need to appeal to introspection, taken as a mode of inner perception, to explain knowledge of inner states. Whatever the merits of that argument, I take 'introspection' to mean whatever method by which we come to know of our inner states. Whether this phenomenon is to be explained along the lines of inner perception, or is a means of 'just knowing' as Shoemaker suggests, is of no importance here.

note that infallibility is not a consequence of self-justification.[8] Taking our original example, the fact that "I am in pain" is self-justifying does not entail that any such judgement is necessarily correct. Even if the judgement is self-justified, it is still possible for someone who is not in pain to believe they are in pain, and for someone in pain not to believe they are in pain. This is because self-justification does not preclude the holding of unwarranted beliefs, and such unwarranted beliefs may be false.

Nevertheless, self-justification does make the occurrence of error somewhat remarkable. This stems from the noted result that a self-justifying judgement can only be false if it is unjustified. Now, whilst people may be generally rational, there is nothing untoward with the occasional lapse of rationality, that is with the formation of an unjustified belief. But with a self-justifying predicate, it is a plausible assumption that part of having the concept is to (tacitly) acknowledge that it is a self-justifying property. That is, it is part of having the concept of pain that you know that the only reason you can have to believe that you are in pain is that you are in pain. In this case, not only do you know what justifies the belief that you are in pain, but you also know if that justification obtains. So the only way that you could form an unjustified (and potentially false) belief would be if you formed a belief for which you had no justification, and for which you knew you had no justification. Error, although possible, can only arise if the subject forms a belief he knows to be without warrant, which is to say he suffers a marked failure in his rationality.

In claiming that the requirements of a rule are self-justifying, we should expect errors in such matters to have the remarkable status noted above. Yet, when following a rule, mistakes are commonplace, and do not have to be explained in terms of a lapse of rationality. For example, I know the rule for addition, and yet certainly can make a mistake when adding large numbers, perhaps because I simply forgot to carry ten at some stage. Similarly, I know the rule governing the use of the word 'red', and yet can sincerely call things 'red' which are not red due to the lighting conditions. In both cases I get the requirements of the rule wrong, and the explanation in each situation is quite mundane. Yet our reason for suggesting that the requirements of a rule are self-justifying is the *prima facie* similarity between the grasp of a

---

[8] The anti-private language argument is directed against the notion that we can have self-justifying beliefs about our inner states (see for example PI §289). My own view is that the traditional anti-private language argument is unsuccessful, and that (following Kripke) the best hope for establishing such a result lies with the rule-following considerations. In any case, our remit is to investigate the scope of the rule-following considerations without recourse to any contentious premises. Therefore we cannot base the argument against rule-following on the anti-private language argument, and so have no reason to dismiss claims of self-justification at this stage.

rule and the holding of a belief. It now appears that judgements about the requirement of a rule and about the identity of ones beliefs have quite different epistemologies, which makes claim that judgements about the application of a rule are self-justifying difficult to sustain.

The difficulty, though, is not insurmountable. For note that in some situations (to be explained below), we can divide the justification for the application of a rule into two parts. That is, the overall judgement:

Application: The rule I am following now requires φ

is justified just in case the following two judgements are justified:

Rule: In situation C the rule I am following requires φ
World: I am now in situation C.[9]

The first is a judgement about the requirement of the rule, the second about the context in which the rule is applied.

The self-justification thesis applies only to the judgement about the requirement of the rule, for whether situation C occurs is a matter of one's physical environment, and so the worldly judgement is empirical, and not self-justifying. Consequently error with respect to the way of the world will be quite unexceptional. Yet error in either judgement is liable to produce an error in the application of a rule,[10] and so the fact that errors in one's appraisal of the world are mundane will serve to explain why misapplication of the rule is also unremarkable: error will usually not be the result of a misapprehension of the requirements of a rule, but of the misclassification of the situation in which the rule is applied.[11] To know what the

---

[9] The fact that rule-following involves the appraisal of two distinct factors - a rule element, and a world element - is highlighted by Wright (1989c).

[10] Error in one or other of the judgements does not guarantee a misapplication of the rule: if both are wrong, then the errors could cancel out.

[11] The possibility of error does arise even without a mis-classification of one's external environment. Taking the example of someone who adds incorrectly, we should note that the rule for addition as usually executed is a composite of several rules: those for aligning numbers, adding columns of single digits, carrying, and so on. The situation under consideration is when I forget to apply one of these rules - the rule for carrying tens, say. The fact that the sum I give is wrong does not then reflect a misreading of the requirements of any one sub-rule, but rather an omission in applying the rule. So the possibility for error does not show that my beliefs about the requirements of each of the composite rules, or the overarching rule for applying the sub-rules, are not self-justified. Since the rule for addition is a composite rule, there is no need to say that one's beliefs about its requirements are self-

requirements of a rule are does not entail that the rule is followed correctly, and so the existence of mundane error does not undermine the self-justification thesis.

It should be recognised that the diagnosis of error in terms of separable judgements is not universally applicable, for some judgements about application are not separable into the two components. For example, in applying the word 'red', we should need an analysis along the following lines:

Rule: The rule I am following requires that I apply 'red' to red things.
World: The object before me is red.

As before, we have a judgement as to the requirement of the rule, and a judgement about my present context in which I am applying the rule. But in this case the concept red appears in the both judgements. Since we have no way of describing the specific contexts in which the word 'red' should be applied other than by using that concept, the account therefore requires that I grasp and apply the concept red, and hence assumes that I know the requirements of the rule in question. But this makes the division into separate judgements circular, and so this device cannot be used to explain the epistemology involved in following such rules.

In situations in which there is no classification of the worldly context which does not require grasp of the target rule, then there is only one judgement that the rule-follower can make, namely:

Application: In my present situation the rule I am following requires that I say 'red'.

Certainly this type of judgement cannot be self-justifying, any more than a judgement about any empirical state of affairs can be self-justifying. Yet there is no means of separating the judgement into two factors, one of which is self-justifying, the other being responsible for any mundane error.

Although the situation cannot be analysed in terms of separable judgements, the same type of diagnosis as was offered above can still be applied in such cases. We merely have to recognise that it is not the role of individual judgements which is important, but rather the

---

justifying, but only that one's beliefs about the requirements of each component rule, together with the requirements of the rule for composition, are self-justifying.

relation between justificatory factors, and truth-makers. All that has been demonstrated is that we cannot always ascribe concepts to our subjects which allow the separation of the various justificatory factors into distinct judgements - namely a judgement based on sensory evidence as to my worldly situation, and a further judgement as to the dictates of the rule. But the fact is that the same two factors - the way of the world and the requirements of the rule - do still feature in my belief-formation process. That is, although no single judgement that I make is justified by the worldly input, and no single judgement is justified by the contribution of the rule, both features in combination justify the single judgement I do make, namely that the rule I am following now requires φ.

Since we do not have separable judgements, it is not possible to say that one reason supports a self-justifying belief, and the other not. Nevertheless, the two inputs do still contribute to the justification of the overall belief in different ways. For not only are there two factors which contribute to the justification of the judgement, but there are also two factors which in combination make the judgement true or false - namely, the way of the world, and the requirement of the rule. The difference between these two factors is that in one case the *same* factor provides both the justificatory element and the truth-making element (that is, the rule is both a justificatory element, and a truth-making element), whilst in the other case the two things are distinct (on the one hand some sensory input will provide the justification, whereas it is the worldly state of affairs which contributes to the truth of the judgement).

In relinquishing a distinct judgement about the requirement of the rule, we cannot continue to refer to the situation as one of self-justification, for there is no judgement which is self-justified. Nevertheless, the point is that with respect to the worldly contribution, fact and justification are distinct, and this is sufficient to explain why mundane errors are possible. And the fact that justification and fact are identical with respect to the contribution of the rule means that it is fallacious to demand that there be any intermediate justificatory element. It is in failing to recognise this situation that Wittgenstein's regress appears to arise. In conclusion, the argument to the exhaustion of justifications fails, and there is no threat to rule-following from this particular epistemological direction.

Platonism

Wittgenstein's thoughts on the epistemology of rules are not wholly unproductive, for the materials at hand do serve to discount a Platonic account of rule-following, and so at least

remove one of the options on the post-Kripke list of meaning-determinants. The argument comes in three stages.

One: under Platonism, judgements about the requirements of a rule are not self-justifying. The distinctive claim of Platonism is that a rule is an abstract object which exists independently of anyone's thought or experience. In that case, it must be possible to be in ignorance of the requirements of the rule, which means that the rule cannot be something which is automatically within anyone's awareness. As a result, there must be some difference between someone who is in epistemic contact with the rule as compared with someone who is not, and this difference must presumably come in terms of a difference in occurrent states of mind. So the norm is a mind-independent abstract object, whereas the reason for belief as to the requirements of a rule involves an occurrent state of mind. So under Platonism, the two things - norm and reason - are distinct, and there is no element of self-justification within the attendant epistemology.

Two: if judgements about the requirements of a rule are not self-justifying, any justification must be based on inference. This follows from the fact that the rule, and one's reasons for judging what the rule requires, are distinct. In that case, when I form a belief as to the requirements of the rule, I form a belief about one thing on the basis of my awareness of something else, a process which must involve an inferential step.

Three: rule-following based on inference is regressive. This is quite straightforward, for the process of inference just is a process of following a rule (rule of inference). So to follow a rule, it is necessary to follow a rule, making the whole endeavour impossible.[12]

The principle claim of Platonism is that a rule is a mind-independent abstract object, and it is precisely this which precludes the possibility of the type of epistemology of rule-following developed above which involves an element of self-justification. Since that epistemology was the only stable means of accounting for rule-following we could identify, Platonism is thereby rendered incoherent.[13]

---

[12] Put differently, the account requires that we interpret some occurrent mental state as evidence as to the requirements of the rule, thus leading us once again to the regress of interpretations.

[13] The argument here bears a degree of similarity to that given by Pears (1988), but there are two important differences. One, Pears suggests that any object in mind could be variously interpreted ("any single thing in anyone's mind would always be connectable with more than one set of applications" (1988, p. 469), but he does not bring out the important distinction between standing in need of interpretation and being susceptible to

Conclusion

In turning from the ontology to the epistemology of rule-following, we have made some important headway. Both the dispositional and the Platonic theories have been discounted. Significantly, though, Wittgenstein's argument that in following a rule we act without reasons has not been substantiated. The only way that a rule can at once determine a truth-condition, and itself contribute to the justification for a belief as to its own requirements, is if the norm is itself constitutive of an occurrent mental state. Only if we reject the existence of a mind-independent object can the norm automatically be 'in mind'. On this picture, there is no object - the rule - which is grasped, but rather a unitary state which we call 'grasp of a rule'. Though this appears to label a relation, it need not. There is no object which is grasped: there is no such thing as a norm which is not 'in mind' - or at least not which informs our behaviour. 'Grasping a rule' is then an esoteric way of saying that someone is in possession of a certain normative state of mind. (So there is no such thing as an ungrasped rule, in the way that there is no such thing as an unexperienced pain.) We have as yet said nothing which serves to discount this *sui generis* theory of rule-following, and although the options have now been somewhat reduced, our search for a case against rule-following *per se* has not yet proved successful.

---

interpretation. (Only the former is regressive, whereas Pears refers to the latter.) The second difference is that Pears finds the argument against Platonism as effective against a non-Platonic view of norms. The argument is, though, very swift, the principle claim being that a non-Platonic norm also succumbs to the regress of interpretations. As I state in the text above, I do not see that a *sui generis* state *needs* to be interpreted to give a norm, nor to inform our behaviour, which is to say I disagree with Pears in this respect.

3. The Indexical Argument

Our search for an argument against rule-following has so far proved unsuccessful: neither Kripke's 'sceptical' argument, nor the epistemological concerns raised by Wittgenstein, can be sustained. In this chapter my aim is to rectify this situation, to present a fresh argument - *the indexical argument* - which does establish the impossibility of rule-following. As we shall see, the indexical argument relies on components generated by the two approaches previously rejected.

One outcome of the foregoing two chapters is to make the target for the indexical argument particularly clear. Both the dispositional and Platonic theories have been rejected, leaving only the possibility that a rule or a meaning be determined by an intrinsically representational, normative mental state. If we are to show that rule-following is impossible, we have to discount this option. In accepting the possibility of infinite norms, we accept that a rule can be in force, that a speaker's word can refer determinately to an infinite number of objects, in an infinite number of situations. As we have seen, we can also give an adequate account for the required connection between the rule and the speaker's action, so that an epistemological challenge to rule-following is no longer viable. So how to proceed?

In order to find a distinctive problem with rule-following, we need to retrace our steps a little. As mentioned in Chapter 1, Kripke sets up the 'sceptical' argument by asking for some fact about the speaker in the past which determines how she ought to act in the present. That is, he portrays the 'sceptical' problem as an issue arising between the past meaning and present behaviour. Yet we also saw that the relation between prior meaning and present action is not instrumental to the 'sceptical' argument, but is only a superficial presentational device. It was because Kripke could find nothing capable of determining an infinite norm at a time that the 'sceptical' argument arose. A standard, once set, can be followed at any time; the fact that nothing determines how I should act in the present merely reflected the supposed failure of anything to determine a norm in the first place.

At least, there is no problem with this last thought - that it is always possible to follow a past standard when in the present - in as much as the rule-following considerations have been developed so far. It is the issue of the trans-temporality of norms which I want to focus on here. Although when speaking I intend to accord with my present meaning, where my present meaning is determined by my present mental state, no doubt the understanding is that what I mean now is the same as what I meant in the past. Indeed, unless meanings have the ability to remain stable throughout time, then language could not function as a means of communication, or as a means of classification. Meaning must remain constant throughout periods of change in the environment if language is to categorise those situations which are themselves subject to change. And if we are to mean the same thing at different times, then we must be able to follow the same rules at different times. If there is no content to the idea of following the same rule at different times, then on that basis alone meaning is not possible - it cannot fulfil its function. In what follows I argue that it is not possible to mean the same thing, to set the same standard, at different times, and that this entails a position identical to Kripke's meaning 'scepticism'.

Representation and Context

To illustrate the problem, we can set up a challenge modelled on Kripke's own example of 'grue'. It will be recalled that 'grue' is defined as follows:

$\forall$x, t: x is grue at time t $\Leftrightarrow$ x is green at t and t < 1st January 2000
x is blue otherwise.

(Note that here the temporally relevant issue is the colour of x as it occurs at t – so that t is the circumstance of evaluation in Kaplan's terminology.) The 'sceptical challenge' was to find some fact which determined that I meant either *green* or *grue* by the term 'green', given that my usage up to now is consistent with either. Our answer to this challenge is that a mental state can determine an infinite norm, and so determines that the term applies to green things in the future, not blue things.

Given our acceptance of intrinsically normative mental states, we can grant that when I utter a word *w* today (in 1999) it refers to all green things, throughout time and space. There is then no question that I do not mean *grue*. However, a problem emerges when we ask what it is to mean the same thing at a later time. Suppose that in 2001 I utter a word *v*, and that *v* refers to all blue objects. Initially we might think that *w* and *v* mean different things, for they

have different extensions. Yet, in the spirit of 'grue', we may readily construct a predicate which puts this is doubt. Let us define *grut* as follows:

$\forall x, t$: ⌜x is grut⌝ is true when uttered at time $t \Leftrightarrow$ x is green and $t \leq 2000$
x is blue otherwise.

(Here t concerns the context of utterance, not the circumstance of evaluation.) The distinctive feature of 'grut' is that it is an indexical predicate. The idea of an indexical expression is clear enough. The referent of "I" "you" "now" "that" etc. depends on some feature of the context in which the expression is uttered: "I" refers to the speaker, "you" to the listener etc. In addition, tensed expressions are also indexical: the extension of "is red" depends upon the time of utterance, for example. Similarly, when used prior to 2000, 'grut' refers to green things; when used *after* 2000 it refers to blue things; so its extension depends on the time at which it is uttered.[1] (To be clear, the claim is that 'grut' refers to green things from the context of the year 2000 no matter when the green things exist – we may use the term to refer to objects in the past, present or future.) This is in distinct contrast to 'grue' which is not indexical: whether "x is grue" is true or not depends only upon the time at which x exists, not at the time the utterance is made.

What 'grut' highlights is that the mere determination of a correctness condition is not sufficient to determine meaning, for the fact that a word refers to all green things is consistent with it meaning either *green*, or *grut*. As well as fixing the extension of a word when used at a time, meaning must also fix the extension of the word when used at different times. On this basis, we may instigate a new challenge: our aim now is to find some property of an individual which determines that she means *green*, not *grut*.

The question can be put on a more general level if we note that every referring term must have (to use Kaplan's 1990, p.37 terminology) a *character*. This is a function from context of utterance to a *content*, where content is a function from possible world to extension. (Given a context of utterance, the character determines the extension for a term in each possible world.) Thus, the character for "I" would be a function determining that the word refers to the person uttering it. Likewise, the character for "in five years time" determines that the expression refers to a time five years in the future of the time of utterance, and so on.

---

[1] The "is" in the above definition is taken to be eternal; it is not the source of any indexicality.

Notably it is not only indexicals which have such functions: a non-indexical has a character which is a constant function, so that its extension is invariant with respect to context of utterance.

In these terms, our question is: what determines the character for a given expression? We have already established that meaning can only be determined by an intrinsically normative mental state - that is a state which represents certain actions in certain contexts as correct. Although in reaching this conclusion the phenomenon of indexicality was not uppermost in our minds, the consideration of indexicality does nothing to immediately invalidate the thesis. To accommodate indexicality, the only new element we need add is that the *sui generis* meaning state must be capable of determining an extension which varies with context.

This point perhaps needs some explanation, if only because with indexicals there is some ambiguity about the nature of meaning. For example, if I say "Today is a momentous day" on Tuesday, and again on Wednesday, in one sense I have said the same thing (at least I uttered the same sentence), whereas in another I said different things (respectively that Tuesday was momentous, and that Wednesday was momentous). That is, the same words express different propositions on the two occasions. (In Kaplan's terminology, they have different contents.)

Despite this ambiguity, it is quite in order to say that indexical expressions each have *a* meaning, a meaning which is constant across contexts. This is because, having been taught a specific language, it is possible for me to understand sentences containing indexical expressions which I have not met previously. In other words, if I know the context of utterance of a sentence containing an indexical, I know the truth conditions of the utterance in question, because I know the meaning. We can therefore safely say that there is a property common to any English speaker who utters "Today is a momentous day" – and a property which remains constant between an individual's utterances at different times – namely the property of knowing the meanings of the constituent words.

So, whether an expression is indexical or not, its meaning must be determined by an intrinsically representational state – call it the *meaning-state*. By definition, to mean the same thing is to be in the same meaning-state. And since character is a part of meaning, the character is something which must be fixed by the meaning-state. Hence, as stated, the

intrinsically representational meaning-state must determine the way in which extension (across possible worlds) varies with context.

Intrinsic Representation and Character

So far, so good, but it is when we come to ask how a meaning-state is to determine a character that problems arise. The obvious way in which a meaning-state may determine an indexical meaning is by being indexical itself. Hence the extension of "now" varies with time because the correctness condition set by the corresponding meaning-state is itself set context-sensitively.

The idea of an intrinsically representative state which is indexical does require an extension of the idea of intrinsic representation. For an entity to be intrinsically representational means, of course, that its representational power is constitutive of its identity. It is quite easy to take this to mean that in order to identify an intrinsically representational state, you have to identify (amongst other things) *what* it represents. This appears to be the thought behind the following, for example:

> for ordinary relations, like *x is sitting on y* or *x is employed by y*, x and y are identifiable entities quite apart from whether they happen to be thus related to each other. The thing which is x would be the same x whether or not it were sitting on y or employed by y. But the same is not true of intentional 'relations'. One and the same belief cannot at one moment be about a frog (that it is green, say) and another moment about a house (that *it* is green). The latter is a different belief. What a belief is supposed to be about is crucial to which belief it is. (Dennett and Haugeland in Gregory 1987, p. 384)

(This passage is specific to beliefs, but the point raised concerns intrinsically representational states in general.)

Dennett and Haugeland's description of the intrinsically representational looks to be too restrictive, for it fails to account for the context-sensitivity of certain representational states. For example, suppose I believed the thing to my left is green, and that there is a frog to my left. My belief was then about a frog - not *explicitly* about the frog of course, but certainly about the frog, for expression "to my left" picks out the frog, and given the way of the world, the truth of my belief depends on the colour of the frog. Suppose further that I still believe that the thing to my left is green, but that I have moved, so that now to my left there is a house. On both occasions I believe (rightly, perhaps) that the thing to my left is green, but whereas that belief was formerly about the frog, it is now about the house.

As discussed with respect to meaning, whether we want to call this the same belief depends on how we taxonomise beliefs (*de dicto* or *de re*). But the point is that no matter how we classify beliefs, it is necessary to accept that there is a representational state which fixes what I am talking about, that this element remains constant between the two cases, and that the state is therefore indexical. Yet by combining the notions of intrinsic representation and indexicality, it is no longer possible to identify representational states by what they represent. Instead it is necessary to identify how what they represent varies with context, which is to say we have to give the character.

Although the idea of an indexical intrinsic representation appears to be straightforward enough, it is actually fundamentally incoherent. The difficulty emerges when we look at the nature of representation in general. To be clear, although the most familiar type of representation is the way that a picture represents an object in virtue of a resemblance between the picture and the thing represented, our concern is somewhat broader, so that representation does not refer only to something involving a resemblance relation, but to the way in which one entity can be 'about' another, or perhaps 'point to' or 'refer to' something other than itself. This, at least is the standard use of the term 'representation' when talking about the representational power of the mind, in the sense that intentional states (beliefs, intentions, etc.) refer to, or represent, things outside themselves.

The feature I am interested in can be identified using two examples of representational objects: firstly, a portrait of the Queen; secondly, an arrow suspended above a row of objects (illustrated below).

⇩
A B C D E

The picture of the Queen represents a particular person - Elizabeth II - and it does so independently of where it is, or when it is (i.e. independently of its spatio-temporal location). If I move the picture to a different room it still represents the Queen; and, should the picture persist into the twenty-second century (even though the Queen undoubtedly will not), the picture would still be a picture of Elizabeth II.[2]

---

[2] Causal relations may enter into the correct account of representation here, so that the picture of the Queen would not be a picture of the Queen if it had been made before the Queen existed. This type of consideration can be ignored in the present case.

This is in distinct contrast to the arrow suspended over the row of letters. Under the broad notion of 'represents', the arrow represents, points to, or refers to, the letter which is directly below it. If above the "A", then it points to the "A", and so on. What the arrow represents is a function of the location of the arrow - move the arrow and you (may) change what it points to.[3] One way of putting this is that the representational power of the arrow is context-sensitive (i.e. with respect to small changes in spatial location), whereas the representational power of the portrait is not.

The example is formulated in terms of physical objects which are not intrinsically representational. In as much as these are physical representations, they have only a *derived* intentionally, a representational power bestowed on them by the mind. Nevertheless, the example is instructive, for the general feature of representation which it serves to identify does not depend on whether the representational power in question is derived or not.

The reason for looking at these examples is to highlight what it means to be indexical. Restricting our attention, in terms of context, to changes in time only, a representation is indexical just in case what is represents varies with time; it is non-indexical just in case what it represents does not vary with time. Crucial to this idea, of course, is the possibility of having the same representation at different times. With physical pictures, the identity conditions of representations are straightforward. If a picture represents in a non-context-sensitive manner, then both the identity of the picture, and the identity of what it represents, are determined solely by its physical form.[4] Given that a portrait of the Queen represents the Queen in a non-context-sensitive way, we know that we have the same picture of the Queen on different occasions just because we know what it is to have an object with the same physical form at different times. When it comes to context-sensitive physical representations, such as the arrow, the situation is similarly uncomplicated. Taking the arrow as the representational object, it is clear that the identity of the arrow does not vary with context (i.e. it is the same arrow which moves over the series of letters). Again the identity of the representational object is given by the identity conditions for physical objects. In both these

---

[3] Actually, the *relative* position of the arrow and letter is what counts - move the letter and you get the same effect - but the point is the same: what the arrow points to depends on the spatial relation between the entities involved.

[4] Here the identity of a picture is taken to be determined by its physical form, so that, for example, two physically distinct copies count as the same picture. The argument can readily be adapted to accommodate the alternative thesis that the identity of a picture is determined by its physical identity.

cases, whether context-sensitive or not, the identity of the representation is given not by what is represented, but by the physical features of the representational object. Although what is represented varies with context, the representational objects have identity conditions that are independent of their (derived) representational powers.

To bring out the problem with indexical representation, consider how issues of indexicality relate to our example of 'grut'. Suppose that yesterday I was in representational state R_1, which determined that the word *w* applies to green objects. This situation is consistent with my meaning either *green* or *grut*, for when uttered yesterday, both terms refer to green objects. (In fact, there are countless predicates we could construct along the lines of grut which are also consistent with my word having this extension at this given time, but for the sake of simplicity we can consider a restricted case between the two given options.) Given that R_1 determines what I meant, then the answer to the question "did I mean *green* or *grut*?" is simple: if the extension of R_1 does not vary with time (i.e. it always refers to green things) then I meant *green*, but if the extension of R_1 changes with time (i.e. it refers to blue things today), then I meant *grut*. My word is indexical just in case the underlying representation is indexical. To pin down what I mean is then simply a matter of identifying the character of R_1.

To do that, we need merely conduct a conceptual experiment: that is, move R_1 to the later time, and see what it refers to from the new time. It is here that the crucial difficulty arises. For to be indexical means that the extension of the representational state varies with time, and so clearly the very idea of (time-) indexicality rests on the possibility of re-identifying a representational state at a later time: it must be possible to have the *same* representational state at different times if it is to make sense to talk of the extension of that state varying (or not) with time. Unless it makes sense to talk of the same representational state at different times, it makes no sense to talk of the extension of that meaning-state varying with time. And if we are to talk about the same representation at different times, there must be something in virtue of which R_1 at t_1 is the same as (or different from) R_2 at t_2.

That is to say, in order to ensure that representational states satisfy the notion of trans-temporal identity, we should hope to identify the property in virtue of which they are the same. In the case of intrinsically representational states, this property is already known. As discussed above, indexical representations cannot be identified by what they represent, for what they represent changes with context. In order to identify an intrinsically

representational state it is necessary to say what the representation represents in different contexts: we have to give its *character*. Hence, the criterion we require is that representations R_1 and R_2 are the same if, and only if, they share the same character.

Is this answer satisfactory, though? Reason to suppose that it is not comes when we note that conflicting requirements have been made of the relation between character and the sameness of representation relation. (For convenience, we can call the sameness of representation relation $same_{rep}$). On the one hand, the essence of the identity condition given above is that it shows how $same_{rep}$ is the product of the respective properties of two representations. Notably, this applies not only to actual instances of the relation, but also to possible instances. In a case where R_1 and R_2 are the same only in some possible world, we should still say that they are the same in that world because they share the same character in that world. Hence, to account for the relation is to show how to *construct* every instance of it.

In addition to this constructional relation between character and $same_{rep}$, there is also a *constitutive* relation to consider. To see this, let us continue to restrict our attention to two times only, namely t_1 and t_2. In that case, to claim that R_1 at t_1 has the character *green* is to claim, roughly speaking, that R_1 represents green things from the context of t_1, and *were* R_1 to exist at t_2, then it would represent green things from then too. To be more accurate, though, we should not refer to what R_1 at t_1 *itself* would represent were it to exist at t_2, for strictly speaking R_1 is a time-slice, and cannot exist at any time other than t_1. In saying that R_1 would represent green things from t_2, we mean to say that in some possible world there is some *other* representation R_2 which exists at t_2, such that R_2 both represents green things and *is the same as* R_1. Hence, possession of a character consists in participation in at least one instance of the sameness relation, albeit in some possible world.

This constitutive relation is at odds with the claim that we have shown how to construct sameness out of character. For, as soon as we have an instance of character on the table, we also have an instance (in some possible world) of $same_{rep}$. Clearly any such instance of the relation is not the product of any construction process – we have simply imported it as a unit, so to speak, in the guise of a property. That is to say, the observation that R_1 has such-and-such a character does not appear to be part of what *accounts* for the fact that R_1 can re-occur; rather, it seems to merely *record* that R_1 can re-occur. In light of this, our account of $same_{rep}$ looks to be defective.

These observations are, admittedly, somewhat impressionistic. They do serve, though, to raise the question of whether the following conditional:

$\forall x \neq y$: x is the same representation as y $\Leftrightarrow$ x has the same character as y

is wholly satisfactory, and hence whether it succeeds in identifying the identity condition on representations.

Having raised this question, it is worth noting that in this respect, appearances can be deceptive: not everything that looks like an account of a relation is one. To see this, consider the following example. Suppose we want to account for a quite different relation – to identify what it is that makes two people mutual sisters. To this end, let us appropriate the term 'sisterhood', and take it to mean a group of females who are mutual (blood) sisters (rather than the usual sense of a group of women bound by a common purpose). In that case, the following conditional is true:

$\forall x \neq y$: x and y are sisters $\Leftrightarrow$ x is in the same sisterhood as y.

On formal grounds, this seems to be an account of what makes two people sisters, in that it is an analysis of the (mutual) *sisters* relation in terms of the properties of the people involved. Nevertheless, it is quite clear that being in the same sisterhood is not what *makes* people mutual sisters in the relevant sense – that is a matter of being female and sharing the same parents.

Clearly there is some similarity between this case and the case of character, but it is probably not worth comparing them directly. The message is simply that, whilst we had thought that to account for a relation we simply had to analyse the relation in terms of the respective properties of the related entities, this is now shown to be insufficient. The example clearly shows that not everything which looks like an account of a relation – not every biconditional with an appropriate form – tells us what we want to know. Given that there is some doubt about the adequacy of the account of $same_{rep}$, and given that the form of the account is no guarantee of success, the obvious question to ask is this: what must an account of a relation tell us in order to be satisfactory? It is only once we have a clearer understanding of this that we can properly decide whether our account of $same_{rep}$ really is deficient.

The key to this issue is the idea that a biconditional may look like an identity condition and yet fail to be an identity condition. How can this be? In general, to identify a condition is to identify what is required to satisfy it – to give its satisfaction condition. The form of a satisfaction condition for condition C is, at its most basic, this:

$\forall x$: x satisfies C $\Leftrightarrow$ Fx.

It is easy to see that not everything that has this form succeeds in identifying a satisfaction condition. For example, suppose we define the property F to be the property *satisfies C*. In this case, the biconditional is certainly true: x satisfies C just in case it has the property of satisfying C. However, it should be clear from the way that F was defined that the claim that x is F tells us only *that* x satisfies C, it does not tell us *why* it satisfies C. Hence this biconditional fails to identify the property required to satisfy C.

The significant feature of the example given above is that, although Fx entails the satisfaction of C, it does not do so *by satisfying* C; it entails the satisfaction of C because F *consists* in the satisfaction of C. The distinction between 'entails by satisfying' and 'entails by constitution' is easy to draw. On the one hand, if Fx entails the satisfaction of C by satisfying it, then it can only yield that satisfaction *in conjunction* with a satisfaction condition. That is to say, the entailment from Fx to satisfaction must depend on the existence of a satisfaction condition. In contrast, if the entailment from Fx to satisfaction does not depend on the existence of a satisfaction condition, then Fx does not produce satisfaction by satisfying the condition – rather, it entails satisfaction directly.

As a result, there is a simple test to tell the difference between a conditional which identifies a satisfaction condition, and one which does not. For, if the given property is what satisfies the condition, the entailment

$\forall x$: Fx $\Rightarrow$ x satisfies C

will depend on the existence of a satisfaction condition. Hence, the assumption that there is no satisfaction condition ought to undermine this conditional. In contrast, if this conditional is not undermined by such an assumption, then the conditional does not depend on the existence of the condition, and the property in question yields satisfaction in some other way.

It may be useful to present this point in another, more direct, way. Suppose we have a biconditional which seems to identify the satisfaction condition for a condition. Of course, the biconditional does not explicitly state that such-and-such is the satisfaction condition, but it still identifies it. The situation is similar to the way the statement "Grass is green" does not state that grass has such-and-such a colour, but it still identifies the colour of grass. Clearly, the assumption that grass has no colour ought to falsify any such statement. Similarly, on the assumption that there is no satisfaction condition for a given condition, any statement that identifies it as having a satisfaction condition ought to come out false. The test of a genuine satisfaction condition is, therefore, to see how it fares under the assumption that there is no satisfaction condition – a genuine condition will be falsified, a sham condition need not be.

Let us apply this test to our putative identity criterion for representations. To do this, suppose that there is *no* identity condition for representations: nothing makes sameness of representation over time possible, and so sameness of representation over time is not possible. That is, there is only representation at a time. Under this assumption, clearly no two representations R_1 at t_1 and R_2 at t_2 (where $t_1 \neq t_2$) are the same. In addition, since no representation can re-occur at another time, no representation can represent anything at (i.e. from the context of) any other time, which means that no representation can have a character. If no representation can have a character, then no two representations can have the same character. (Perhaps it is better to say that under our assumption of no identity condition, no representation can have a *total* character. Without the possibility of re-identification over time, a representation-at-a-time can only represent from the one time at which it exists, with the value of the function for all other times remaining undefined. Hence, it will only have a *partial* character, that is, a partial function from the time it exists to whatever it represents from that time. Clearly, though, it remains true that no two representations at different times can have the same character, partial or otherwise.)

We have, then, on the basis of our assumption, the following two results: (1) that no two representations at different times are the same; (2) that no two representations share the same character. This is to say that each side of out putative identity condition, which is this:

$\forall x \neq y$: x is the same representation as y $\Leftrightarrow$ x has the same character as y,

is false. Consequently, the biconditional overall is true. On the assumption that there is no identity condition for representations, our putative identity condition remains true. As the

biconditional is not falsified by the assumption that there is no identity condition, it itself cannot be the identity condition. Or, to put the point the other way, although the combination of characters entail that two representations are the same, this entailment does not depend on the existence of an identity condition, and they do not yield sameness by satisfying the identity condition. Either way, we have to conclude that that biconditional does not identify an identity criterion at all.[5]

If character is not what satisfies the identity criterion, we should like to claim that some other property performs this role. However, the conditions any such property should meet are incompatible. As we have seen, to determine meaning it is necessary to determine character; any property that determines character must entail the possibility of re-identification; and any property which entails the possibility of re-identification is subject to the argument given above.[6] Hence, whatever satisfies the identity condition for representational states must determine character, and yet by the argument given above cannot determine character, which is to say there can be no such identity condition.

---

[5] By way of contrast, consider what happens in a similar situation to an 'acceptable' trans-temporal identity condition. One candidate for the identity condition for persons over time is this:

$\forall x \neq y$: x is (a segment of) the same person as y $\Leftrightarrow$ x has the same soul as y.

To follow the course of the reasoning used above, on the assumption that there is no identity criterion for persons over time, it certainly follows that no two persons-at-a-time are the same, so the LHS of the biconditional is false. However, we have no reason to deny that two people may have the same soul, for the possession (and identity) of a soul does not depend on personal identity over time. Any claim we should make about the possession of a soul is unaltered by our assumption that personal identity is impossible. Hence, we have as much reason as we ever did to accept that two segments of a person might share the same soul, and so no grounds on which to claim that the RHS of the biconditional is false. As far as we can tell, then, the LHS is false, yet the RHS could still be true, and hence the biconditional overall seems false. As we would expect, this legitimate example of prospective identity condition is rendered false by the assumption that there is no identity condition.

[6] One characteristic of the exposition is that it proceeds very much on an extensional level: sameness of meaning is considered to consist in sameness of character, where character is entirely extensional. Yet we have good reason to believe that more than mere extension-determination is involved in meaning ("2+2" and "4" refer to the same thing in all possible worlds, but they have different meanings). It is for this reason that we need a notion of sense, and it is the sense of an indexical expression which we should expect to remain constant even though its extension changes. Could we not find identity conditions on senses, and use that to decide whether any two representational states are the same?

The short answer is: no. The idea is that sameness of representational state is determined by sameness of sense, and that sense determines extension across contexts – that is, it determines character – but not *vice versa*. However, (to follow the reasoning used above) on the assumption that there is no identity condition on meaning-states, it follows that no meaning-state can re-occur, and so no two states can have the same character (sameness of character requires the possibility of re-identification, as above). Since sameness of sense determines sameness of character, it follows (by MTT) that no two states can have the same sense. As a result, the claim that sameness of sense is equivalent to sameness of meaning is true (i.e. each is impossible) even under the assumption that there is no identity condition on meaning, and so sameness of sense would not constitute an identity criterion on representations.

Since nothing can account for the re-identification of meaning, meaning the same thing at different times is impossible. To put the point in terms of our example, suppose that yesterday my utterance *w* refers to green things, and that today it refers to all blue things. Then, insofar as we consider only the extensions of these utterances, it is quite possible that on both occasions I meant "grut". It is equally possible, though, that yesterday I meant green, and today I mean blue, and that I have changed what I meant. The issue is decided by the identity of the meaning-state – has that changed or not? – and to that question there is no answer. Nothing determines whether this is the same, or a different state: the matter is underdetermined.

A Rule-Following Consideration?

The immediate conclusion of the foregoing argument is that nothing determines whether I mean *green* as opposed to *grut*. The destructive power of this result is brought out when two additional facts are borne in mind. One is that indexicality is not a peripheral linguistic phenomenon: many of the statements we utter include some indexical component (including tensed expressions, many names, pronouns, and demonstratives).[7] In light of the pervasiveness, indexicality has to be recognised as a central feature of our language, one which any theory of meaning has to accommodate.

The other additional consideration is that temporal indexicality is but one type of indexicality. The extension of an expression may also vary with the place of utterance ("here", "five miles away"), the identity of the speaker ("I", "my granny"), the identity of the audience ("you", "your best friend"), and so on. Indeed, the extension of an expression may vary with any alteration in *any* feature of the context of utterance: *anything* is a potential 'index' for an indexical. If meaning is determined by a property of an individual, and the individual is in some context, then every element of that context is *potentially* relevant to the determination of the extension. However, nothing can determine which elements of the speaker's context are relevant (if any). Change the context in any way, change the extension of the expression in any way, and you could have the same meaning on both occasions, simply by construing the expression as indexical with respect to the relevant change in context. Therefore, whenever there is a change in the context of the speaker (and that means

---

[7] The pervasiveness of indexicality is emphasised by Barwise and Perry (1983) and Searle (1979). Whilst in some cases an indexicality can be removed if it is replaced with an explicit description, in many others it cannot, a point made admirably by Perry (1979).

with every passing second), there is no fact of the matter as to whether he means the same thing as he did in the previous context. But that means that the idea of meaning the same thing in different contexts is empty, and therefore so is the idea that *sui generis* mental states determine meanings.

It is quite evident that the example of 'grut' borrows heavily from Kripke's use of 'grue'; but the substance of the argument is significantly different. Kripke, following Wittgenstein, notes that any series of utterances may accord with some rule or meaning. Therefore a meaning must pick out one series of actions, and determine that just those utterances are correct. That means the meaning must discriminate an infinite number of contexts (for an action is classified in terms of the context in which it is performed). Kripke wonders how any mental state can do this - how can 'green' refer to an infinite number of (potential or actual) green objects, in any number of different contexts? The answer is of course that the meaning picks out all green objects in virtue of their greenness. The representational power of a mental state can determine that the colour of the object is the only relevant feature, and that any other context is irrelevant. If we accept the intrinsic representational power of the mind, then there no reason to think that meanings are underdetermined in the way that Kripke proposes. In contrast, in the argument given here, I do not consider the context of the objects being referred to, but the context of the speaker. Given that in context X the speaker refers (potentially) to an infinite number of objects, we want to know what determines whether the fact that the expression has this extension is dependent on the context of the utterance. Given that sometimes the extension of a meaning is dependent upon context, what makes it so? It is only when we have this that we have what it is that determines meaning, and the argument above suggests that nothing can perform this task.

Why call this an attack on rule-following? Despite the obvious affinity to Kripke's interpretation, is it not rather a direct attack on meaning, akin more to the work of Quine than Wittgenstein? Whilst there are several passages in the *Investigations* in which Wittgenstein raises issues concerning indexicality which have some bearing on rule-following (see for example PI §163 and §226), it is fair to say that Wittgenstein fails to draw any substantial message out of them. Significantly, though, it is a mistake to limit the problem of rule-following, as Kripke does, to the problem of determining an infinite correctness condition. A rule must do more than determine a correctness condition. It must determine a correctness condition which remains stable as the environment changes. It is implicit in the idea of rule-following - which is after all an on-going activity - that the same rule governs one's

behaviour over a period of time. The result of the indexical argument, although formulated in terms of meaning, can equally be applied to rules. Any two correctness conditions, grasped in different contexts, may be the same rule. If yesterday my calling green things 'green' accorded with the rule, and today calling blue things does, then it is possible that on both occasions I am following the rule for *grut*. It appears that what is needed is a further rule, one which dictates how the correctness condition of the rule I am following varies, or does not vary, in different contexts. But then of course the same problem arises for that rule. Yesterday this second-order rule may have dictated that the extension of my first-order rule remains the same over time, today it may dictate that it change in a grut-like way, and yet be the same rule. As before, if one rule does not fix a correctness condition across all variations in context, there is no use in looking to other rules to do that job. That is why rule-following is impossible.[8]

---

[8] How does the example of the arrow fare under the indexical argument? As noted above, an arrow's representational power is derived from the representational power of the mind. An arrow 'represents' in the way it does because that is how we think of it (i.e. mentally represent it) as representing. In order to have this derived representational ability, it is necessary for us to be able to think of the arrow's representational power in the same way over a period of time. That is, the bestowing representation of the arrow's representational power must remain the same over time. Since such re-identification of a representational state is precisely what has proved to be impossible, such derived indexical representation as displayed by the arrow is also impossible.

It might be thought that this loss of our paradigm case of indexicality has an adverse affect on the status of the argument. After all, the argument uses the paradigm to characterise representation – if the paradigm turns out not to exemplify the phenomenon we are interested in, how can it fulfil its illuminating function? Looking at the nature of the argument, though, this worry is unfounded. The purpose behind the arrow example was to identify a key feature required for indexical representation (namely possession of a trans-temporal identity condition), before showing that this requirement cannot be satisfied in the (more fundamental) mental domain. The fact that the apparently straightforward indexicality of the arrow is undermined in the process does not falsify the claim that a trans-temporal identity condition is necessary for indexical representation. In the execution of the argument, then, the 'paradigm' case is of interest not because it actually is an example of indexical representation, but rather because it (initially) looks like a paradigm case. Even though the reality proves to be different, this initial appearance can still be used to identify the necessary requirements for indexical representation.

PART TWO

RELINQUISHING REALISM

4. Meaning Irrealism

The position now established is exactly that of Kripke's 'sceptical paradox': no property of a speaker can determine which rule she is following, no fact about her determines what she means. The threatened conclusion is that no one is a rule-follower, and that no one means anything when they speak. Our task is to avoid this untenable result.

The problem of meaning nihilism arises because we have been following a course parallel to Kripke's: both the 'sceptical' and the indexical arguments conclude that rules are radically underdetermined, and it is such underdetermination which puts pressure on meaning. For his part, Kripke goes on to offer a 'sceptical' solution to his 'sceptical' problem, a position designed to avoid meaning nihilism. In outline, the strategy is this: by giving an *irrealist* account of rules and meaning (in a sense to be explained below), Kripke claims that there need be no 'fact of the matter' in order for us legitimately to say that someone means something. As a result, we retain the right to ascribe meanings to people as we normally would, despite the underdetermination identified by the 'sceptical' argument, and consequently the 'sceptical' argument does not give us the right to *deny* that someone means something if we would normally say that they do. Given that there are many situations in which we do indeed say that people make meaningful utterances, the 'sceptical' argument does not licence the conclusion that no one means anything, and so does not entail meaning nihilism.

Given the similarities between the indexical and 'sceptical' arguments, it is worth asking whether the same type of strategy may be deployed with respect to the indexical argument. The motivation is the same, namely that of avoiding meaning nihilism, and given that the net result of irrealism is that our ordinary assertoric practices are to be respected, the 'sceptical' solution should operate with equal effectiveness against *any* kind of underdetermination claim. If irrealism was a suitable response to Kripke's argument for nihilism, then it ought to be appropriate here too.

As we shall see, the initial idea here is right: the 'sceptical' solution is suitably indiscriminate; meaning irrealism is *as* effective as an antidote to the indexical argument as

it is to the 'sceptical' argument. This, though, raises the questions: how effective is the solution in the first place? Providing an answer will be the central concern of this chapter.

However, before offering a critical appraisal of the irrealist position, work has to be done to give a characterisation of the 'sceptical' solution with adequate precision. As with the exposition of the 'sceptical' problem, Kripke's presentation of the 'sceptical' solution, though generally clear, has spawned various differing interpretations. To assess whether meaning irrealism could contain the destructive effects of the indexical argument, it is essential we know exactly what the thesis is, and how it operates.

Kripke's 'Sceptical Solution'

Kripke makes three basic claims under the banner of the 'sceptical' solution. The first is a rejection of truth-conditions in favour of assertion-conditions:

> All that is needed to legitimise assertions that someone means something is that there be roughly specifiable circumstances under which they are legitimately assertable. (Kripke 1982, p. 78)

The second is the inclusion of utility considerations:

> Wittgenstein's general picture of language...requires for an account of a type of utterance not merely that we say under what conditions an utterance of that type can be made, but also what role and utility in our lives can be ascribed to the practice of making this type of utterance under such conditions. (Kripke 1982, p. 92)

And the third is:

> It turns out that this role, and these conditions, involve reference to a community. (Kripke 1982, p. 79)

The picture we get is that we are warranted in saying that "Jane means plus by '+'" just in case Jane agrees, for the most part, with the rest of the community members when asked to do additions. Importantly, Kripke is *not* saying that agreement with the community is a necessary and sufficient condition for rule-following. Nor is it the consensus which determines correctness, that whatever answer the community gives to an addition is necessarily correct. Rather, agreement with the community merely licences the assertion that the speaker is following a rule; that in making an ascription we dignify the subject as a member of the linguistic community; and these assertion-conditions and the utility of this practice give the content of "Jane means plus".

In this way, meaning nihilism is avoided. So long as the subject satisfies the assertion-conditions (so long as she continues to agree with the community), and so long as the practice of making such utterances serves some purpose, meanings may be legitimately ascribed. The ascription of content does not then depend on there being any 'fact of the matter'; and consequently the practice of making such ascriptions is not undermined if it is shown that there is no 'fact of the matter'. We can thus continue to say that people mean things as we did before.

Different interpretations of the 'sceptical' solution have arisen because Kripke's critics have given each of the three elements Kripke mentions (assertion-conditions, utility, community) varying degrees of emphasis, and various different roles within the solution. To quickly reduce the number of possible interpretations, we can at this stage remove all mention of the community from consideration. Although Kripke's thesis if often billed as a 'community account', and although the community is of undoubted philosophical importance,[1] it is a mistake to see the community as *instrumental* to the workings of the 'sceptical conclusion'. This is because the community is only invoked when we come to *identify* the particular assertion-conditions, and the utility, that meaning ascriptions have. Quite what licenses the assertion that Jane means plus, and quite what role it plays in our lives, are important issues when it comes to assessing the ramifications of the 'sceptical' solution, but for the purposes of avoiding meaning nihilism (which is the task in hand) all that matters is that such utterances can legitimately be made. On the assumption that assertion-conditions and utility considerations can between them accomplish this, then all that is needed for the solution to be effective is that 'Jane means plus' has determinate assertion-conditions and/or use. To discern whether the 'sceptical solution' can, in principle, solve the 'paradox of rule-following', it is thus unnecessary to identify the particular assertion-conditions/utility considerations involved, and so for the time being the community aspect can be ignored.[2]

With respect to the remaining two elements, Kripke's insistence that *both* assertion-conditions *and* use are essential to the theory is puzzling, for each appears in its own right to

---

[1] The role of the community is of particular significance for the relationship between the rule-following considerations and the anti-private language argument (see Kripke 1982, pp. 98-105).

[2] There is something of a consensus that communal aspect of the 'sceptical' solution is misguided. (See McGinn (1984a p. 185), Boghossian (1989b, pp. 520-522), Goldfarb (1985, p. 483) and Blackburn (1984b, pp. 293-295).) My own view is that neither side establishes their respective view satisfactorily, but the point is that even if Kripke is wrong, the remaining elements which capture the essence of the 'sceptical' solution remain intact. The issue of how the community is involved in the warranted ascription of a rule is taken up in detail in Chapter 7.

be satisfactory as a 'sceptical solution'. To see this, suppose we first drop all mention of utility, and hold simply that statements such as "Jane means plus" are assertable when the relevant assertion-conditions are satisfied. In doing this, the fundamental benefit of the 'solution' is retained, in that that the right to make a meaning ascription does not depend on evidence for the truth of the claim, and so the meaning can be attributed irrespective of any 'fact of the matter'. In that case, the 'sceptical' result that nothing makes such statements true does nothing to undermine our right to say that Jane means plus.

On the assumption that assertion conditions are at least an element of the meaning of "Jane means plus", it is not clear how the addition of utility considerations can improve the theory. In defence of a utility element, Kripke says, "Such a role must exist if this aspect of the language game is not to be idle." (1982, p. 75). Certainly we can accept that if no such role existed, no one would have any reason to make the assertion in question, but what is not clear is why the use should be dignified as being constitutive of the meaning. The point is brought out when we compare the relationship between meaning and use when the meaning of an expression is given in terms of truth-conditions. We can use the sentence, "Wittgenstein was born in 1889" for many purposes - to express a belief, to get someone to do something, to tell a joke, to give an example, etc. - but in each case the *meaning* of the expression is the same. Indeed, it is because we know what the statement means, and because we know that our audience shares this knowledge, that we can use the statement in these various ways. If this is right, the utility of a statement depends upon its meaning, but is not part of that meaning.

This situation does not appear to be altered if truth-conditions are replaced with assertion-conditions. In order to use an expression, a shared grasp of its meaning is crucial, but whether the meaning is to be explained in terms of truth-conditions rather than assertion-conditions is - in this respect at least - not germane. The fact that a given expression has communally grasped assertion-conditions would enable us to use that expression for many different purposes; again it would be *because* the expression has the meaning it has that we can use it to influence other people in the way we do. Thus, even when truth-conditions are replaced with assertion-conditions, it appears to be unnecessary to include a description of use *within* the account of the meaning.

The same type of observation applies when we consider an alternative interpretation of the 'sceptical' solution which occurs when we remove all mention of assertion-conditions to

leave a theory of meaning given only in terms of utility considerations. Just as an assertion-conditional account does not need utility considerations, so too an account which mentions utility has no call for assertion-conditions. To see this, though, the sense in which 'utility' is used has to be explored in more depth.

Kripke exemplifies the type of use that statements about meanings have as follows:

> the utility of this practice can be brought out by considering...a man who buys something at the grocer's. The customer, when he deals with the grocer and asks for five apples, expects the grocer to count as he does, not according to some bizarre, non-standard rule; and so, if his dealings with the grocer involve a computation, such as '68 + 57', he expects the grocer's response to agree with his own....Our entire lives depend upon countless such interactions, and on the 'game' of attributing to others the mastery of certain concepts or rules, *thereby showing that we expect them to behave as we do*. (Emphasis added. Kripke 1982, pp. 92-93)

> Even though, when we play this game and attribute concepts to individuals, we depict no special 'state' of their minds, *we do something of importance. We take them provisionally into the community*, as long as further deviant behaviour does not exclude them. (Emphasis added. Kripke 1982, p. 95)

In light of these comments, Kripke has been interpreted as offering an expressivist account of meaning ascriptions.[3] Expressivism (emotivism) is most familiar as an ethical theory, according to which saying "x is good" is not to describe the world, but merely to express a certain attitude of approval towards x. When applied to semantics, the expressive theory holds that to say "Jane means plus" is not to describe Jane in any way, but just to express a certain attitude about her. Hence, the purpose of such statements serves to dignify them as a member of the linguistic community, to express an attitude of acceptance, and of expectation that they should continue to accord with others in their behaviour.

Expressivism about meanings does offer sufficient resources to assuage the paradox of rule-following. If the discourse of rules and meaning merely serves for the expression of attitudes, then such discourse is not answerable to any fact. Despite the indexical argument, we do still distinguish between those initiated into our linguistic community and those not, between those whose use of English we expect to follow established patterns. Under the expressive theory, it remains quite proper to express these attitudes by uttering statements such as "Jane means plus". Indeed, to say "Jane does not mean plus" would be to express an attitude of exclusion, and of a lack of expectation as to Jane's future behaviour: attitudes

---

[3] The expressivist interpretation of the 'sceptical' solution is given by Wright (1984), Boghossian (1989b), Blackburn (1984b), and Heal (1989).

which, despite the indexical argument, we do not have. So under expressivism the statement that meaning is impossible (that no one means anything) is certainly not licensed.

If the expression of an attitude is constitutive of the meaning of meaning ascriptions, then there is no obvious role within the solution for Kripke's other component - assertion-conditions - to play. This is because it is the very nature of expressive discourse that it does not answer to objective states of affairs, but only to one's subjective reactions. To take a typical situation, we may suppose that A has received a normal induction into the English language, and uses it in accordance with the rest of the community. On the basis of A's past performance, B accepts him as an English speaker, and expects him to continue to use the language in the normal way. To express this, B says "A means *noodle* by 'noodle'", and the like. In this case we may be tempted to say that A's behaviour is the basis for B's assertion, and that such behaviour features within the assertion-conditions for B's statement. Yet if "A means *noodle*" is to express an attitude, it does not matter *why* B has the attitude in question. No matter how A has behaved in the past, if B feels that he is a member of the linguistic community, then expressing this attitude by saying "B means noodle" is an appropriate thing to do. In other words, the expression of an attitude exhausts the linguistic function of the statement, leaving assertion-conditions no room in which to make their own semantic contribution.

To consolidate, there are three positions before us: the 'sceptical' solution is interpreted as a theory based on (a) assertion-conditions alone, (b) expressivism alone, or (c) a combination of the two. So far, we have found a certain difficulty in making sense of (c), if only on the grounds of parsimony. Otherwise no one option presents itself as preferable to the others.

Happily, there is not need to decide between the three options, for it turns out that they are all versions of the same theory. To see this, consider the class of expressions which undoubtedly do have expressive meaning, namely interjections ("Blast!", "Alas!" "Eureka!" and so on). Certainly such utterances do not serve to describe the world in any way, and yet it is still possible to use them inappropriately. For example, were I to hit my thumb with a hammer and shout "Hooray!", I would not be using the term in accordance with its meaning. The fact that there is such a thing as using an expressive in accordance with its meaning demonstrates that expressive utterances are governed by appropriate norms.

For an utterance to express a certain attitude can only mean that the sentence is only uttered with full propriety by someone who has the attitude in question. Transferring this result from interjections to indicative sentences with expressive content, in saying that a sentence such as "Apple pie is tasty" expresses a certain attitude towards apple pie, what we mean is that the assertion is correctly made only by someone who has the attitude in question. By this token, we may say that possession of the appropriate attitude *licenses* the assertion, making the attitude the assertion-condition for the given sentence. In that case, the key difference between what have so far been labelled 'expressive' and 'assertion-conditional' theories lies only with the type of situation which licenses the utterance. For an orthodox assertion-conditional statement the assertion-conditions concern objective, worldly states of affairs, whereas for an expressive the norm involves the affective states of the speaker (attitude, mood, or emotion). Seen from this perspective, expressivism is actually a *version* of the assertion-conditional theory.[4]

Applying this result directly to meanings, we still have three available theses, differing only in terms of the identity of the assertion-conditions governing meaning ascriptions. To take a specific example, my saying "Johnny means plus" may be warranted by: (a) Johnny's past behaviour; (b) my attitude towards Johnny; or (c) a combination of the two. Under the first thesis it is only Johnny's past performance which licenses my assertion that he means plus, whether I am confident that we will agree in the future or not. In contrast, the second thesis holds that Johnny's past performance is not *semantically* relevant to my utterance; although I may have the attitude in question because of Johnny's past performance, it is the attitude itself (no matter how it arose) which warrants the assertion. In this case it is my attitude which is semantically relevant, whilst Johnny's behaviour is not.

It should now be clear that nothing precludes a theory which adopts both of the elements mentioned above (which is the closest we can get to the two-factor theory forwarded by Kripke), for there is no reason why the assertion-condition for a statement be satisfied only by a combination of factors. In the situation at hand, "Johnny means plus" would be assertable only if Johnny has behaved in an appropriate manner in the past, *and* I expect him

---

[4] Horwich (1990, pp. 87-88) also suggests that expressive meaning can be considered a type of assertion-conditional meaning.

to continue to behave in a specified way in the future, so that the utterance is warranted by both objective and affective elements.[5]

It is unnecessary at this stage to assess the relative merits of our three theories any further. The significant point is that the essence of the 'sceptical' solution is captured wholly in terms of autonomous assertion-conditions. (Henceforth I shall take 'meaning irrealism' to refer to the assertion-conditional theory.) It is the existence of assertion-conditions not dependent upon truth-conditions which, it is claimed, stymies meaning nihilism. The actual identity of these assertion-conditions is not the immediate concern; what does matter is whether the underlying principle is sound, whether the adoption of an assertion-conditional theory of any kind can have the desired result.

Assessment of the Irrealist Solution

To assess the viability of meaning irrealism, I shall concentrate on one question: is meaning irrealism coherent?[6] To this end, there are two main charges to consider. One is that meaning irrealism does not have the promised placatory effect, that it itself is susceptible to the indexical argument. The other is that meaning irrealism is incoherent on quite different grounds. Let us consider each in turn.

The most obvious reason to suspect that irrealism is inadequate as a response to the indexical argument is that the theory replaces one norm (truth) with another norm (warranted assertability). Since the replacement is motivated by the result that following a rule is impossible this appears to be little in the way of an improvement: the theory appeals to the very notion - grasp of a norm - which the indexical argument attacks. Following this line of reasoning through, irrealism states that the content of "Jane means plus" is given in terms of assertion-conditions, so that an understanding of this sentence is given in terms of one's

---

[5] Although he does not advertise it as such, Blackburn (1984b) appears to propose such a 'conjunctive' theory. Blackburn accepts the negative force of Kripke's 'sceptical' argument, and suggests that the correct response is expressivism about meanings. He says: "In my view [the sceptical argument] invites a projectivist [expressivist] explanation of these kinds of judgements....In any event, *we are left searching for standards whereby to make that judgement.*" (Emphasis added, Blackburn 1984b, p. 300). So Blackburn thinks that there must be *standards* which govern use *in addition* to the presence (or absence) of the attitude in question, which is to combine the two elements under discussion.

[6] To keep the scope of this discussion within reasonable limits, I shall assume that assertion-conditions are *semantically* adequate; that is that they are in general capable of giving an adequate account of linguistic content. The matter is contentions - see for example Appiah (1986) and Kirkham (1992) - but the focus of our discussion in not assertion-conditional semantics overall, but only the specific application of such a theory to rules and meaning.

grasp of a norm of warranted assertability. Yet on the basis of the indexical argument, nothing determines which norm I grasp - and that includes the norm of warranted assertability. So if the meaning of "Jane means plus" is assertion-conditional, it follows that no one ever grasps the meaning of that expression. Consequently it would not follow that anyone has the right to ascribe meanings to others, and the solution would collapse.

This objection, however, ignores the very power of the irrealist position. If norms are assertion-conditional, then a norm can be ascribed to anyone who satisfies the specified assertion-conditions. Therefore, by the light of the theory itself, we can be warranted in ascribing grasp of a norm of warranted assertability to a subject *even though* the indexical argument shows that the matter is strictly underdetermined. It is thus *assertable* that someone grasps the norm of warranted assertability that is necessary to say that Jane means plus. So the irrealist solution does no more than appeal to irreal norms, and so, does not succumb to the indexical argument as suggested.[7]

A second concern is whether meaning ascriptions do actually have any satisfiable assertion-conditions. The issue originates with a criticism directed at Kripke by several authors. Baker and Hacker (1984, p. 37), McGinn (1984a, p. 188) and Wright (1984, p. 770) all state that even if meaning ascriptions are assertion-conditional, we still cannot draw an appropriate distinction between someone who means *green* and someone who means *grue*. For, if in the past Jane has called all green things 'green', then since here behaviour is consistent with her meaning either *green* or *grue*, the claim that Jane means *grue* is *as warranted* as the claim that she means *green*. The point applies with equal force against meaning irrealism proposed in response to the indexical argument. In the same way, someone who means *green* and someone who means *grut* ought to behave in the same way up until a certain point in time (the point at which the extensions diverge), and so until that time there would be nothing they could do which would distinguish between them. So it is always assertable that Jane means green, *and* that she means grue, *and* that she means grut, and also any number of other such predicates. But then we lose grip on the idea that a definite meaning can be ascribed to a given individual: there are an infinite number of equally warranted candidates, and so which description ought to be used is as radically underdetermined as it ever was.

---

[7] This objection to meaning irrealism is mentioned by McGinn (1984, p.183-184) (who attributes it to Field), and rejected for the reason as given above.

Again, though, the objection fails to recognise an essential feature of the irrealist thesis. Usually, we expect that the assertion-conditions for a statement are those states of affairs which count as good evidence that the statement is true. To identify a statements assertion-conditions, we need merely consider what would make it true, and thus discern what would indicate the satisfaction of this truth-condition. In short, assertion-conditions depend upon truth-conditions.

In contrast, under irrealism, there are no truth-conditions to play this determining role; rather, any assertion-conditions are autonomous, and constitutive of meaning. Consequently, it is not possible to reason from truth to assertability. The best we can do in order to identify the assertion-conditions for a given statement is to apply our knowledge as users of the language; after all, the claim is that knowledge of meaning consists in knowledge of assertion-conditions, so we should all have tacit knowledge of what those conditions are. To this end, it is significant that even once we are familiar with the grue-like alternatives, we do still accept that people mean green (and not grue, or grut). This in itself is a good indication that such behaviour satisfies the assertion-conditions for green and nothing else. There may be no rationale behind this assertoric practice, but the point is that none is required.

Boghossian on Meaning Irrealism and Truth

With these objections out of the way, we can accept that meaning irrealism is not itself susceptible to the indexical argument. The remaining question is whether meaning irrealism is internally coherent. The first argument I want to consider which suggests that it is not is given by Boghossian, who attempts to exploit the connection between meaning and truth to find a weakness in the irrealist thesis.

Boghossian (1989, 1990a) offers a *reductio ad absurdum* on meaning irrealism. His strategy is to show that meaning irrealism entails that truth is both a property and also that it is not a property, this being a contradiction. The argument starts with a negative characterisation of meaning irrealism: rather than stating that meaning ascriptions have assertion-conditions, Boghossian gives irrealism as the denial that they have truth-conditions. That is:[8]

---

[8] The presentation of the argument is slightly different between Boghossian's (1990a) and (1989). Although Boghossian describes (1990a) as being the more detailed of the two, the structure of the account given here follows (1989) more closely, simply because I find that version clearer. I have not followed the numbering system Boghossian uses to identify the steps in the argument.

(1) For all S, p: ⌜S means that p⌝ is not truth-conditional.

Here p is a propositional variable, and S ranges over all significant declarative sentences. "Significant" and "declarative" means that

> the sentence possesses a role within the language: its use must be appropriately disciplined by norms of correct utterance; and that it possesses an appropriate syntax: it must admit of coherent embedding within negation, the conditional, and other connectives, and within contexts of propositional attitudes. (Boghossian 1990a, p. 163)

As mentioned, Boghossian's aim is to show that meaning irrealism entails that truth is both a property and not a property. Theories of truth which claim that truth is not a genuine property may be termed 'deflationary' theories, as opposed to 'robust' theories. If truth is not a property, then the truth 'predicate' must have some grammatical role other than that of assigning a genuine property to a truth-bearer. That is, although claims such as "S is true" and "It is true that p" *appear* to ascribe the property truth to a sentence S and a proposition respectively, the deflationist holds that the surface grammar here is misleading. Thus, for example, on the redundancy theory, to say that "grass is green" is true is just to say that grass is green; and, on the expressive theory, to ascribe truth to a sentence is to pay it some kind of compliment, to signal one's endorsement of it. In contrast, robust theories state that the surface grammars of "It is true that p" and "S is true" are not misleading, that such sentences do attribute a property to a truth-bearer. The various robust accounts then differ in the analysis offered of that property (coherence, correspondence, ideal justification, and so on).

Without restricting himself to any particular version of deflationism, Boghossian claims that if truth is deflationary then *any* assertoric sentence must be truth-conditional:

> Any meaningful, declarative sentence would be (at minimum) a candidate for an assertion; it would be, thereby, a *candidate* for the compliment we pay sentences we are prepared to assert....Any such sentence would count, therefore, as truth-conditional in a deflationary sense. (Boghossian 1990a, p. 165)

This is plausible enough. If the truth predicate is just a device for disquotation - for cancelling out the effect of quotation marks - or serves simply as a means of endorsing an assertion, then anyone asserting S will have as much right to say that S is true. In other words, there can be no grounds on which it can be denied that a declarative sentence is truth-conditional simply in virtue of its subject matter; the only relevant features are syntactic. That is:

(2) If truth is deflationary, for all S: ⌜S⌝ is truth-conditional.

However, the initial premise - meaning irrealism - states that meaning ascriptions are not truth-conditional. That is, (1) is the claim that some sentences do not have truth-conditions. Thus we have:

(3) For some S: ⌜S⌝ is not truth-conditional.

Statement (3) is the negation of the consequent of (2), and so by MTT on (2) and (3) we get:

(4) Truth is robust.

If truth-conditions are to be denied of certain sentences on the grounds that they have a certain subject matter, then truth cannot be merely a device for disquotation, or for the endorsement of those statements we are prepared to assert. Meaning irrealism is incompatible with deflationary truth.

The next step is to show that meaning irrealism is also incompatible with robust truth, by giving a subsidiary *reductio*. First, Boghossian notes that:

> Since the truth-condition of any sentence S is (in part, anyway) a function of its meaning, a non-factualism about meaning will enjoin a non-factualism about truth-conditions: what truth-condition S possesses could hardly be a factual matter if that in virtue of which it has a particular truth-condition is not itself a factual matter. (Boghossian 1989b, p. 524)

This seems right. It is precisely because nothing can determine a truth-condition that meaning irrealism is required in the first place, so in as much as truth-conditions can be ascribed to a sentence, we should expect it to be under the auspices of irrealism. As a result, (1) entails:

(5) For all S, p: ⌜S has truth-condition p⌝ is not truth-conditional.

Boghossian continues:

> Judgements about whether an object possesses a robust property could hardly fail to be factual. If P is some genuinely robust property, then it is hard to see how there could fail to be a fact of the matter about whether an object has P. It does not matter if P is subjective or otherwise dependent upon our responses. So long as it is a genuine, language independent property, judgements about it will have to be factual, will have to be possessed of robust truth conditions. In particular, if truth is

> a robust property, then judgements about a sentence's truth value must themselves be factual. (Boghossian 1989, p. 526)

So given that truth is robust (from (4)), we get:

(6) For all S: ⌜S is true⌝ is truth-conditional.

This is the statement to be contradicted for the subsidiary *reductio*. To establish its negation, note that just as the truth-conditions of a sentence are a function of its meaning (premise (2)), so too the truth-*value* of a sentence is a function of its truth-conditions. Yet, as Boghossian says:

> There is no way...that a sentence's possessing a truth-value could be a thoroughly factual matter...if there is a non-factuality about one of its determinants. (Boghossian 1990a, p. 175)

Therefore, if the truth-conditions of a sentence are recorded by a statement which is less than factual, then so too the truth-value of the sentence must be recorded by a statement which has less than factual content. That is:

(7) For all S, p: ⌜S has truth-condition p⌝ is not truth-conditional ⇒ For all S: ⌜S is true⌝ is not truth-conditional.

The antecedent of (7) is (5), and so by MPP on (5) and (7) we get:

(8) For all S: ⌜S is true⌝ is not truth-conditional.

Statements (6) and (8) are contradictory. This contradiction rests on premises of meaning irrealism (i.e. (1)) and robust truth (i.e. (4)). On the assumption of meaning irrealism, it follows that:

(9) Truth is not robust.

So meaning irrealism demands that truth is both robust and deflationary, which is why Boghossian claims that meaning irrealism is false.

Boghossian's Truth Platitudes

There is something of a tradition of supporting meaning irrealism against this argument by objecting to one or other of Boghossian's assumptions about the nature of truth. For example Kraut (1993) suggests that a deflationary theory of truth does not entail that *every* declarative sentence is truth-conditional. He says: "It is no part of *deflationism as such* that 'true' expresses a compliment we pay all sentences we are prepared to assert; a more selective, less promiscuous compliment might be involved (Kraut 1993, p. 257). In that case, on the assumption that truth is deflationary, it does not follow that every significant declarative sentence is truth-conditional (in short Kraut rejects (2)). Devitt and Rey (1991) criticise Boghossian for assuming that there must be some notion of truth applicable to some sentences (1991, pp. 95-97). They suggest that it is possible that *no* sentences are true or false - a position they call "austere eliminativism". Thus Boghossian is wrong to assume that truth is either robust or deflationary - it might be that there is no such thing as truth at all. Wright (1992, Appendix) in turn thinks that it is a mistake to hold that only *one* notion of truth applies across all discourses. Instead, it could be that different notions of truth apply in different situations - so that the truth of a mathematical statement may be different from the truth of an ethical statement. If so, then it could be that a deflationary notion of truth applies to sentences about semantic issues, with a more robust notion in operation elsewhere. In that case, the premise that truth is either robust or deflationary is again false, for it could be both.

Whatever the merits of these views about truth are, I shall not rely on any one of them as a means of refuting Boghossian. The assumptions Boghossian makes about the nature of truth are quite orthodox, and certainly occupy the default position. Unless meaning irrealism itself can be used to directly motivate the rejection of any one of them - and there is nothing of this ilk in sight - then some additional argumentation is required to motivate this type of rebuttal, which would take this discussion outside its remit.[9] (Certainly we ought not re-characterise truth simply to prop up the irrealist theory.) Instead, I prefer to accept Boghossian's assumptions, and show that the argument fails on its own terms.

---

[9] I am not suggesting that such motivations are not forthcoming. Wright (1992), in particular, does give reasons in support of his claim that truth is many-sorted.

Irrealism and Deflationary Truth

Boghossian's argument is unsound because meaning irrealism is actually compatible with deflationary truth. To isolate Boghossian's error, let us reconstruct the argument as follows. From the premise:

(2) If truth is deflationary, for all S: ⌜S⌝ is truth-conditional.

and the logical truth:

(*) For all S: ⌜S⌝ is truth-conditional ⇒ For all S, p: ⌜S means that p⌝ is truth-conditional.

we get the following the conditional by transitivity:

(**) Truth is deflationary ⇒ For all S, p: ⌜S means that p⌝ is truth-conditional

Boghossian then takes irrealism:

(1) For all S, p: ⌜S means that p⌝ is not truth-conditional.

to be the contrary of the consequent of (**), giving the result that truth is robust by MTT.

The core claim is that the consequent of (**), namely:

For all S, p: ⌜S means that p⌝ is truth-conditional.

is the contrary of irrealism, namely:

(1) For all S, p: ⌜S means that p⌝ is not truth-conditional.

But is this really the case?

In fact it is not, for there is an equivocation here over what it is to be truth-conditional. This comes out when we look at the motivation for irrealism. The aim of irrealism is of course to give a competing account to the truth-conditional theory of what it is to know the meaning of

statements such as "Jane means plus". Saying that this sentence is assertion-conditional is to make a constitutive claim: meaning consists in, and is to be explained in terms of, assertion-conditions. Similarly, the claim that "Jane means plus" is *not* truth-conditional is the claim that the meaning of this sentence does not consist in, and is not to be explained in terms of, truth-conditions.

In contrast, merely stating that S *possesses* truth-conditions makes no such constitutive or explanatory claim. To say that S *has* a truth-condition is not to say that the meaning of S *consists* in that truth-condition. (Indeed, if possession of a truth-condition is a trivial consequence of possession of assertoric content, we *cannot* explain content in terms of truth-conditions on pain of circularity.) *Contra* Boghossian, it is therefore quite consistent, under an assertion-conditional account of meaning coupled with a deflationary theory of truth, to say both that S *has* the truth condition that p, and yet that S *is not* truth-conditional.

The apparent contradiction is the product of an inappropriate nomenclature, for under the deflationary theory, meaning is not to be explained in terms of truth, and so there are no such things as 'truth-conditions' in the constitutive sense. What is meant is that any sentence which is a candidate for assertion is also a candidate for truth, entailing only that there are circumstances in which the sentence may be called 'true'. A more fitting term for the possession of such deflationary 'truth-conditions' would be 'truth-apt'. Unlike truth-conditionality, talk of truth-aptness is not a constitutive claim, and fulfils no explanatory role. If truth is deflationary, every significant declarative sentence is truth-apt, but *no* sentence is truth-conditional. It does not follow from the claim that meaning ascriptions are not truth-conditional that truth cannot be deflationary: sentences can lack truth-conditions and yet remain truth-apt. As a result, there is no inconsistency of the type identified by Boghossian between meaning irrealism and deflationary truth, and the *reductio* fails.

Meaning Irrealism and Explanatory Power

The message of the foregoing discussion is not wholly negative, for the manner in which Boghossian's argument fails gives us something to build on. As noted, the above rebuttal of Boghossian's argument exploits the distinction between two senses of 'truth-conditional': one constitutive and explanatory, the other non-constitutive and non-explanatory. As it happens, meaning irrealism is eventually untenable because it both rests on this distinction, and yet is unable to maintain a stable position on either side of it.

The relation between irrealism and explanation comes out most directly in the present context when we consider an explanation of a rule-follower's behaviour. Looking first at the realist position, grasp of a rule is an essential element of a standard *psychological explanation*. In many cases, we think people act as they do because they are following certain rules. For instance, the fact that someone expands the series 2, 4, 6,...is to be explained in terms of their intentions, together with the fact that they grasp the rule *add 2*. When ascribing rules to others, we do so in order to explain their behaviour, and so grasp of a rule is inferred on the basis of inference to the best explanation.

Irrealism is motivated by the demonstration that nothing can explain the behaviour of a (so-called) rule-follower in the way that a rule should explain it. (Since it is impossible to follow a rule, the explanation given above falls down.) As a result, a statement about a rule cannot be warranted on the basis of inference to the best explanation, but must be made on some other grounds. For the irrealist, the replacement justification is provided at the semantic level: it is part of the meaning of the expression that the ascription of the rule is warranted in such-and-such circumstances.[10]

Somewhat more significantly for our purposes, in addition to the *descriptive explanation* mentioned above, irrealism also removes the power of rules to feature in what we may call a *prescriptive explanation*. This type of explanation, rather than accounting for why someone behaved in such-and-such manner, is an explanation of why someone *ought* to so act. The form of this explanation will be as follows: to explain why S ought to φ, it need only be mentioned that (a) S is following rule R, (b) R requires φ in C, and (c) S is in situation C.

To see why irrealism annuls this type of explanation we need to consider the consequences irrealism has for truth as identified above. The result of our treatment of Boghossian's argument is that meaning irrealism enjoins a deflationary theory of truth. (The strand of his argument showing that irrealism is incompatible with robust truth still stands.) That is, just as we may legitimately say things such as:

S means that p

we may also just as legitimately say both:

---

[10] Kripke notes (1982, p. 97) that meaning irrealism involves a loss of explanatory power.

"S means that p" is true

and:

"S means that p" has a truth-condition.

Whilst Boghossian construes this result as being a result about the trivial applicability of the truth-predicate, what amounts to the same thing is the claim that, since such sentences are assertable, they have assertion-conditions and are assertion-conditional, not truth-conditional. (Basically the trivial applicability of truth means that the assertion-conditions for "'S means that p' is true" are the same as the assertion-conditions for "S means that p".) Hence:

"'S means that p' has a truth-condition" is assertion-conditional.

Since we can at least say that "S means that p" has a truth-condition, it seems reasonable that, when claiming that someone knows the meaning of "S means that p", we can also claim that they know what the truth-conditions for this sentence are.

Significantly, although we may *say* that knowing the meaning of "S means that p" is to know its truth-conditions, such a description cannot be used to support a prescriptive explanation. For otherwise we could argue that since (1) Bill knows the meaning of "S means that p", and (2) this means he knows what makes it true, it follows that (3) he ought to say "S means that p" only when he has reason to believe that the truth-condition is satisfied. But as we know from the indexical argument the truth-condition for "S means that p" can never be satisfied. Therefore (5) Bill (on familiarisation with the indexical argument) ought never say "S means that p".

It is precisely this type of result which meaning irrealism is designed to overcome. To this end it is essential that although we may *say*:

Bill knows the truth-condition (truth-rule) for "S means that p"

this description cannot be used in a justification - or, equally, to demonstrate the *lack* of justification - for Bill's actions. Hence the fact that:

"'S means that p' is truth-conditional" is assertion-conditional

must entail that the claim:

"'S means that p' is truth-conditional"

cannot be used in a prescriptive explanation.

Clearly there is nothing specific to the ascription of truth-conditions here: *any* statement ascribing grasp of a norm (truth-conditions, assertion-conditions etc.) which is itself assertion-conditional cannot support a prescriptive explanation. This general point gives us our first premise:

(1) $\forall$S, S is assertion-conditional $\Rightarrow$ S is non-explanatory

where "non-explanatory" relates to prescriptive explanation.

In a similar vein, it is significant that the statement of irrealism must itself support prescriptive explanations. This is because, in replacing realism, the irrealist thesis must explain why, *despite* the indexical argument, it is *still correct* to ascribe meaning to people, and to thereby explain how meaning nihilism is avoided. Of course, it does this by providing an alternative explanation of our assertoric practices. In particular, in saying:

$\forall$S, p: ⌜S means that p⌝ is assertion-conditional

the intention is to offer an explanation for why we *ought* still to ascribe meanings to people despite the fact that the matter is strictly underdetermined. Such a prescriptive explanation depends upon the fact that irrealism is a claim about what it is to know the meaning of a sentence, and this claim is about the psychological make-up of whoever knows the meaning of, say, "Jane means plus". Unless meaning irrealism constitutes an alternative explanation of what it is to know the meaning of "Jane means plus", it cannot explain why someone who understands this sentence is nevertheless warranted in asserting it. Significantly, faced with

an argument that meanings are never ascribable, it is necessary to give some reason - an explanation - for why this direct conclusion is mistaken. Without such an explanation, irrealism becomes merely that flat denial of the conclusion to a sound argument, reducing the irrealist thesis to an exercise in stone-walling. Our second premise marks this essential feature of the irrealist thesis:

(2) The claim that ∀S, p: ⌜S means that p⌝ is assertion-conditional is explanatory.

The final element in the argument has already been mentioned. As noted above, meaning irrealism must itself be irrealist: the thesis is a claim about the norms which give the meaning of "Jane means plus", and, by its own lights, it can only be *assertable* that "Jane means plus" has such-and-such assertion-conditions. To make the claim more formally, note that assertion-conditions are norms, and hence are subject to the indexical argument. That is to say, on the assumption of meaning irrealism (i.e. (1)), in as much as we can ascribe assertion-conditions, we can only do so on an irrealist basis. Hence:

(3) ∀S, C: ⌜S has assertion-condition C⌝ is assertion-conditional.

If it is only ever assertable that S has a specified assertion-condition, it can at most be assertable that S is assertion-conditional. Hence:

(4) ∀S: ⌜S is assertion-conditional⌝ is assertion-conditional.

We can of course substitute ⌜S means that p⌝ for S in (4) to give:

(5) ∀S, p: ⌜⌜S means that p⌝ is assertion-conditional⌝ is assertion-conditional.

Similarly, substituting ⌜⌜S means that p⌝ is assertion-conditional⌝ for S in (1) gives:

(6) ∀S: ⌜⌜S means that p⌝ is assertion-conditional⌝ is assertion-conditional
⇒ ⌜⌜S means that p⌝ is assertion-conditional⌝ is non-explanatory.

MPP on (5) and (6) gives:

(7) ∀S, p: ⌜⌜S means that p⌝ is assertion-conditional⌝ is non-explanatory,

which contradicts (2). This is a *reductio* on meaning irrealism.[11]

To put the argument succinctly, the incoherence arises from the following three propositions:

(i) Meaning irrealism is an explanatory thesis.
(ii) Irrealism strips its subject matter of its explanatory power.
(iii) Meaning irrealism is self-applicable.

It is the self-applicability of meaning irrealism - a property peculiar to irrealism about content - which makes it unstable. As a statement about meanings, meaning irrealism applies to itself; it therefore strips itself of all explanatory power. And yet the very value of the thesis rests with its capacity to explain. Meaning irrealism must be explanatory, yet by its own lights it cannot be explanatory, and this makes the position self-defeating.

---

[11] Wright (1984, p. 770), Heal (1989, p. 165) and Boghossian (1989, p. 524) all suggest that meaning irrealism globalises, making all discourse irreal. Wright and Heal find this result to be destructive of irrealism, on the basis that it makes the statement of irrealism itself irreal, but as Boghossian indicates (1989, p.525) there is no evident incoherence in this. (Wright's view on the matter has subsequently changed; see his 1991, appendix.) The above discussion gives one way in which the argument may be developed. Under irrealism, to identify the meaning of a statement, we should look at the conditions under which speakers find the statement to be assertable. The significant point is that no argument (such as the indexical argument) could alter the fact that S is assertable in just those circumstances. If in fact every statement is assertion-conditional, then plausibly this process could be extended to cover the statement of semantic *realism*. In particular, it is normal practice to say that "S means that p" is truth-conditional, and that this statement about truth-conditions is itself an explanatory claim. Under global irrealism, we should never have grounds to revise this practice, which would mean that we could not deny that "S means that p" is truth-conditional, nor indeed that "'S means that p' has truth-condition q" is an *explanatory* claim. The indexical argument would have no power to overturn the licence we have to describe something as an explanatory statement, any more than it can overturn our right to ascribe meanings. In that case, we could not deny the thesis of meaning realism *qua* explanatory claim, nor assert the thesis of meaning irrealism. Under globalisation, meaning irrealism would not be assertable, which would rule it out of contention. I prefer not to rely on this approach, for it is questionable whether meaning irrealism does indeed globalise.

5. Creating Rules

The prospect of meaning nihilism is still with us. The second strategy I want to consider for avoiding this disastrous conclusion is catalysed by Wittgenstein:

> "What you are saying, then, comes to this: a new insight - intuition - is needed at every step to carry out the order '+n' correctly."....It would almost be more correct to say, not that an intuition was needed at every stage, but that a new decision was needed at every stage. (Wittgenstein, *Philosophical Investigations*, §186)

If we were to take Wittgenstein's suggestion without qualification, so that following a rule is a matter of making a new decision at every stage, then we would be likening the expansion of a series to the improvisation of a melody. At any moment in time, it is up to the musician to decide, note by note, what comes next. The musician is continually at the 'creative threshold': those notes already played are fixed elements of the tune, but beyond this nothing is determined. Applying this picture to the expansion of the rule 'add 2' - what *I* mean by the rule 'add 2' - we get a picture in which, as I expand the series beyond the limits I have so far reached, the correct answer is only fixed as and when I make a judgement on the matter. Taking our cue from the model of the improviser, we thus think of my verdict as having the logical status of a decision, fixing each element at the threshold of enquiry, with respect to those elements of the series not considered previously. My verdicts extend the rule into virgin territory, as it were. To give an alternative picture, we should not consider a rule as a fixed rail, pre-formed, available to be followed; but rather the track is laid, continually, at one's feet.

It is, though, no accident that Wittgenstein is more cautious: he says only that talk of a decision would be "*almost* more correct". The modifier ("almost") is required because in 'following a rule', one does not *choose* from a number of alternatives (cf. PI §219); and if our verdicts as to the requirements of a rule do play a determining role, we are not necessarily aware of this fact. (In contrast, a decision is typically known to be a decision.) Instead, one takes the *only* course of action which seems to be correct, with the belief one is responding to an objective, predetermined standard. From a logical point of view, though, the difference between actual decision and verdicts which are merely akin to decisions is minimal. The central idea Wittgenstein here signals is the shift from a rule-*follower*,

someone who acts in the hope of according with a pre-ordained pattern, to rule-*maker*, someone who has a hand in the fixing the requirements of the rule in an on-going fashion.[1]

How can this type of picture help deal with the indexical argument? It will be recalled that the indexical argument shows that, given any two correctness conditions in different contexts, nothing determines whether they are the same correctness condition or not. A direct response to this argument would be to identify some factor which could determine the trans-contextual identity of rules in a way not previously considered. It is in this capacity that the Wittgensteinian suggestion has some bearing. For, to return the picture of musical improvisation, the musician's decision must not only determine which particular note comes next, but his decision must also determine that this next note belongs with the preceding ones, that it is part of the last tune, and not the beginning of a new one. In deciding what comes next, the improviser decides what counts as the *continuation* of the activity already in progress. If there is no more to what comes next than what the musician decides is next, then there is no more to being the same tune than that is what the musician decides is the same tune.

Applying this model to rules, if my on-going decisions determine what is the correct application of a rule in a given situation, then they must also determine that I am following the same rule as before. Whilst the indexical argument shows that nothing about me *as I am now* determines what I have to do *in the future* in order to follow the same rule, the current proposal is that future correctness is not something which has to be determined now, it only has to be determined when the time comes. Whether I am following the same rule as before has the logical status of a decision, and so we should not expect there to be a 'fact of the matter' as to whether I am following the same rule as before until my verdict has been made. In this way, we should not expect anything to antecedently determine what it is to follow the same rule over time; this is a matter to be fixed as we go along.[2]

---

[1] An alternative formulation: "A rule is not an extension. To follow a rule means *to form an extension* according to a 'general' expression." (Emphasis added, Wittgenstein Ms 165 c. 1941-44, 78 - unpublished. Malcolm's translation. Quoted in Malcolm 1989, p. 8).

[2] In keeping with the presentation of the indexical argument I here concentrate on the need for trans-temporal identity conditions for rules. On-going determination would solve the problem of trans-temporal identity, but the indexical argument shows that there are no identity conditions across *any* context, not just the temporal. It is plausible, thought not certain, that whatever factors enter into the determination of trans-temporal identity could also supply the more general requirements for trans-contextual identity, but this can only be ascertained when the determining factors have been properly identified. For the moment I treat on-going (temporal) determination as a test case.

To be clear, the basic idea we are interested in is that certain subject matters may be *judgement-dependent*: that whether a certain situation obtains depends on whether a suitably placed agent thinks it obtains. In its simplest form, this means that for agent S the following conditional holds:

S judges that p $\Rightarrow$ p

with the additional requirement that it is *because* S judges that p that p. (As I say, this is the simplest form; we shall examine a more sophisticated account below). In this respect there are two types of judgement we are interested in. To counter the indexical argument, all that is required is that the following be true, with a left-to-right order of determination:

(I) S thinks she is following the same rule as before $\Rightarrow$ S is following the same rule as before.

Call this the *identity thesis*. In contrast, the more directly Wittgensteinian thesis is that the following conditional be true, with a left-to-right order of determination:

(II) S thinks that the rule she is following now requires φ $\Rightarrow$ the rule S is following now requires φ.

Call this the *application thesis*. If this latter conditional is true, though, then so too must the prior one be true, for as discussed, if the subsequent requirements of a rule are to be fixed as we go along, so too must the fact that we are following the same rule.[3] The judgement-dependence of application (conditional II) is sufficient for the judgement-dependence of (trans-contextual) identity (conditional I), which in turn is sufficient to answer the indexical argument. The question is: can either type of judgement-dependence be endorsed?

---

[3] If this type of picture is to be of any use in dealing with the indexical argument, then the model of an improvisation, or rule prolongation, has to be handled properly. In particular, the predicates 'grue' and 'grut' highlight that it is not just with greater numbers or unobserved objects that we encounter 'new cases, but that at *every* step, with each passing second, any judgement we make concerns an as-yet unencountered situation. Consequently the fact that *yesterday* I continued as series 1004, 1006,... does not determine that *today* I should do the same. We should not, therefore, think of the on-going determination as securing anything which has relevance for future applications. *Each time* I come to consider what follows 1002, the matter is not fixed on this occasion until my verdict is in.

On-Going Determination and Content

The most obvious difficulty with the application thesis is that it appears to entail that whatever seems right is thereby right, a position famously rejected by Wittgenstein (PI § 258). Whatever the logical difficulties with such an identification may be, it certainly fails to account for our ordinary linguistic practices. In talking about the world, errors are possible - what I think is right is not always right - and so can be discarded on the grounds of inadequacy.

However, such an extreme picture is not the only option. All that is required for on-going determination (and hence for a response to the indexical argument) is that the verdict of the 'rule-follower' make a *contribution* to the correctness condition of the rule. The only point that need be established is that what is correct at each step is not *wholly independent* of the subject's judgement on the matter, that there is at least an *element* of 'decision' involved in determining what is the right answer.[4] So there is no reason to suppose that judgement-dependence cannot accommodate the possibility of error, thus disarming the objection.

A second concern is that in embracing any form of on-going determination, we lose grip on the idea that our words have fixed extensions. If, when John says "Apples are red" the term 'red' does not have a fixed extension - if nothing yet determines what 'red' applies to in the future - how can we say that 'red' has a meaning at all? Is it not precisely because 'red' has the (fixed) extension that it has its meaning?

In answer to this, it need only be noted that whilst a theory of meaning which encompasses some notion of on-going determination would be a radical, wholesale, revision of our concept of language (in essence the incorporation of a subjective element to all subject matters), the fact that is it a radical overhaul is not sufficient reason for us to reject it. Whilst it is always a difficulty to say when a proposed revision exceeds the elastic limit of a concept - is this an acceptable revision of F, or have we started talking about something else? - we cannot reject a (potential) revision just because it is a revision.[5]

---

[4] Pears (1988, p. 465) emphasises the fact that for Wittgenstein the rule-follower makes a contribution to, but does not wholly settle, the requirements of a rule.

[5] Sullivan states that "trivially, a rule is not something one can make up as one goes along" (1994, p. 162), so that something which does not determine future applications is not a rule. We should not here get caught up in issues of nomenclature: whether we should try to maintain that on-going determination is still worthy of the name "rule-following", or should instead be viewed as a replacement notion, is of little importance. What does matter is

Rather than arguing in isolation whether content can withstand the type of revision suggested, a more productive method is to look at the motivation for any such a position. (Showing that content *must* involve on-going determination would be the best means of answering anyone who claims that such a revision is untenable.) Notably, though, the indexical argument, on its own, does not properly motivate a thesis of on-going determination. When presented with any argument establishing some kind of underdetermination, it is always a potential response to say that the matter is subjective, that the missing determinant is one's own view on the matter. But in the abstract, such a solution is entirely *ad hoc*. Unless there are independent reasons to suppose that anyone's opinion - and in particular the subject's own opinion - should have any bearing on the matter, then the introduction of judgement-dependence is entirely arbitrary. We cannot appeal to judgement to play a determining role simply because we have a determination deficit: it has to be established that judgement is a determining factor which has hitherto gone unrecognised. The question, therefore, is whether, in light of the indexical argument, there is any reason to suppose that meaning is subjective in either of the senses suggested.

Wright on the Epistemology of Intention

The most promising independent argument for the on-going determination of rules is given, over the course of a series of papers, by Wright (1987, 1988, 1989a, 1989b, 1989c).[6] The route taken is somewhat elliptical, for Wright's initial focus concerns, not rules, but the nature of intention. It is, though, the similarities between rules and intentions which leads Wright to examine intentions in this light in the first place, and so it is no surprise that his thoughts about intention in turn have bearing on rules and meaning.

In Chapter 1 we noted that Wright has been a major proponent of the idea that Kripke's 'sceptical' argument fails to show that meaning is not *sui generis*. Whereas Kripke is resistant to the possibility that a mental state could determine a norm over an infinite

---

whether the theory of on-going determination - whether we describe this as rule-*following* or not - can support an adequate notion of content.

[6] The lion's share of Wright considerable contribution to the discussion of rule-following has been concerned with some version of on-going determination. In his (1980) he attributes to Wittgenstein an argument, based on the anti-realist premise that meaning must be manifestable, to the conclusion that (on-going) communal agreement determines the correct application of a word. (The communal theory is one version of a dispositional theory. As demonstrated in Chapter 2, no dispositional theory can avoid the charge of arbitrary reduction of correctness.) In a later paper (1986) he attempts to establish the same result but without recourse to the contentious anti-realist premise. The series of papers referred to in the text offer a more sophisticated argument, the central difference as regards the conclusion being that all mention of the community is now missing.

domain, Wright, in contrast, hopes to make the *sui generis* response quite palatable by noting that the supposedly troublesome characteristics of meaning - infiniteness and normativity - are shared by common-or-garden intentions. An intention is normative in that it determines a satisfaction condition - some actions satisfy my intention to mow the lawn, others do not - and so the intention provides a standard against which action may be measured. In addition, an intention may be, in the relevant sense, infinite: if I intend to call every red thing I come across "red", then the intention determines that, from an infinite stock of possible actions in an endless number of different situations, an infinite subset of these actions satisfy the intention (i.e. calling red things "red"). Of course, the nature of intention - and particularly the possibility of a naturalistic reduction - is a subject of contention, but the idea that intentions are *sui generis* is at least familiar, and the onus is certainly on anyone who thinks otherwise to make a case. In light of the similitude between meaning and intention, it is as plausible that meaning is a *sui generis* mental state as it is that intending is *sui generis*. Hence *sui generis* meanings should be taken as the default position, and accepted unless shown to be untenable. Initially, then, it is because of the similarities between rules and intentions that Wright is persuaded that rules are immune from Kripke's 'sceptical' attack, and that there is nothing manifestly wrong with the conservative, *sui generis*, account of rule-following.

Having suggested that intentions illustrate the point that infinite normativity is an acceptable property of mental states, Wright subsequently argues that the idea of infinite normativity, as exemplified by intentions, is not so straightforward after all. The problem arises when we try to account for the epistemology of intention, and in particular when trying to give a satisfactory account of the way we know our own intentions. Wright finds that the only way to accommodate the acknowledged first-person epistemology of intention is to accept that intentions are judgement-dependent: that the identity of one's own intentions depends (in part) on what one believes them to be.

This result is significant in the present context when it is applied to what we may call *prolific intentions* - intentions such as the intention to add, which is satisfied only by an on-going series of actions. (In contrast the intention, say, to put the cat out is satisfied by a single act.) There are two ways that the result of judgement-dependence may be applied to prolific intentions. The obvious way is for my intention to add to depend on my belief that I intend to add. The second way arises from the fact that to identify an intention is to identify what satisfies it. Given that a prolific intention is satisfied by a sequence of actions, it may be that

what satisfied the intention is fixed in a piecemeal fashion. So rather then my judgement about the overall identity of the intention fixing all its requirements in one go, I may make a sequence of judgements about the sequence of actions which satisfy the intention. In this way, the requirements of the intention is subject to on-going determination.

The conclusion that intentions may be determined in an on-going manner has a bearing on rules and meaning in two distinct ways. Most directly, Wright's thesis, as restricted to intentions, is in itself enough to establish that rule-following is not an enterprise governed by objective standards. The intention to follow a rule is a paradigmatic prolific intention, and so, if the above result stands up, will be subject to on-going determination. Since the identity of the intention is fixed as we go along, that can only be because the identity of the rule being followed is likewise fixed as we go along. If intentions are judgement-dependent, then quite apart from the objectivity or otherwise of rules themselves, there can be no pre-determined fact as to which rule I am following, and so no objective rule-following.[7]

The alternative strategy is to note that rules and intentions are sufficiently alike that the argument developed in terms of intentions may be applied afresh to rules. That is, if rules share the same troublesome first-person epistemology of intention, it might be that the only means of accommodating their epistemology is to adopt (in addition) judgement-dependence with respect to rules. And just as the judgement dependence of prolific intentions gives rise to on-going determination, so too might the judgement-dependence of rules give rise to the on-going determination of rules. In this second case, then, it is on the basis of an analogy that we arrive at the desired conclusion.

In practical terms both routes reach the same place: the intention to follow a rule does not fix a correctness condition in advance. Nevertheless, we should respect the fact that there are two distinct theses here (which are both endorsed by Wright). In one case the argument for judgement-dependence is applied to intentions, in the other to rules. Although following similar paths, the arguments are somewhat independent. Rules and intentions are analogous to a degree - and this is the very reason why intentions are considered in the first place - but there are differences between them, and so it is possible that the core argument is sound when applied to one, but not when applied to the other. In order to assess Wright's thesis, therefore, it is necessary to consider each of the two options on its own merits. The argument

---

[7] Wright adopts this type of argument (1987, pp. 402-403).

is originally formulated with respect to intentions, and so we should consider that strand of the argument first.[8]

Intention, Introspection, and First-Person Authority

With respect to the epistemology of intention, it is commonplace that we each carry a certain first-person authority about the contents of our own minds. That is, we can all be presumed to know what our own intentions are, and to be able to articulate them as and when we choose. Notably, the ground for my self-knowledge is quite different from that available to other parties. Whereas others can only infer the identity of my intentions on the basis of my overt behaviour, I need have no recourse to such publicly available evidence, nor make no use of any kind of inferential practices. I know my own mind in virtue of it being mine.

The standard explanation for this type of self-awareness is in terms of introspection. Introspection is considered analogous to sense-perception, and is considered a kind of 'inner sight', the essential difference between introspection and sense-perception being that introspection is directed inwards, not outwards. As a result, I have a perspective from which to appraise the contents of my own mind directly, a perspective which no one else shares. The authority of avowal is, on this account, the product of a privileged mode of access.

Wright argues that the idea that we know our own intentions by introspection rests on a conceptual confusion. He develops this thought using specific examples from the writings of Wittgenstein - someone coming to understand the meaning of a word, continuing a series, whistling a tune, recalling what they were going to say, and deciding to play chess.[9] To consider one of these, suppose that I instruct you to continue the series 2, 4, 6, etc. When you get to 1000, 1002, I certainly know that you are doing just what I intended you to do - that is I know that saying "1002" after "1000" is in accordance with my initial intention. The standard explanation for this is that I know by introspection what accords with my intentions. However, when I asked you to continue the series 2, 4, 6, etc. I did not - indeed could not - *consciously* entertain all the (countless) elements of the series. That is, I did not consciously

---

[8] Wright's argument readily falls into the basic structure I have used to characterise the rule-following considerations (negative argument/positive proposal). Since Wright's negative thesis exploits difficulties which face the very idea of rule-following from an epistemological point of view, it could have been examined within Chapter 2. For structural reasons (mainly to avoid giving separate treatments of his negative and positive theses) I have preferred to delay all consideration of Wright's argument, but as consideration of the negative argument is a piece of 'unfinished business', the treatment here is perhaps fuller than it might otherwise be.

[9] See Wright (1987, pp. 396-399).

identify each of the particular series of actions which you would have to carry out in order to satisfy my intention. But then, as Wright says:

> How...can my authority for the claim that at the so-and-so manyth place I intend you to write down thus-and-such, be based on introspection; if, as has been stressed, nothing which went on in me and which has any plausible claim to be regarded as a state of consciousness, explicitly anticipated the so-and-so manyth place at all? (Wright 1989b, p. 396)

The problem is not so much the grounds on which I claim that such-and-such is the next step in the series when I come to consider it. Rather, the point is that to identify an intention is to identify what satisfies it. If I do not run through all the elements of the series in advance, then I cannot be in a position - at least not on the basis of introspection - to say that I intended this one function over some quus-like alternative. Since it is not possible to introspect what my intention requires at every step, introspection cannot reveal the identity of the intention. So how is first-person authority possible?

Judgement-Dependence and the Provisional Conditional

Wright takes this argument to show that the idea that first-person authority stems from acts of inner cognition - from a special kind of epistemic achievement - is false. To replace this picture he says:

> So far as I can see, there is only one possible broad direction...to take. The authority which our self-ascriptions of meaning, intention, and decision assume is not based on any kind of cognitive advantage, expertise or achievement. Rather it is, as it were, a *concession*, unofficially granted to anyone whom one takes seriously as a rational subject. It is so to speak, such a subject's right to declare what he intends, what he intended, and what satisfies his intentions; and his possession of this right consists in the conferral upon such declarations, other things being equal, of a *constitutive* rather than a descriptive role. (Wright 1989b, pp. 400-401)

In saying that the declaration has a constitutive, not descriptive, role, Wright signals the shift to judgement-dependence: I intend to φ because I say/believe that I do. Although we have yet to fill in the detail of the account, it is not difficult to see how judgement-dependence solves the problem of first-person knowledge. Wright himself compares the situation to that of a tennis umpire (1987, p. 401), whose word determines whether a given ball was (officially) in or out. (The umpire is, of course, supposed to respond to the actual physical location of the ball, but the point is that for the purposes of the game it is the umpire's verdict that counts.) In this case there is absolutely no mystery as to how the umpire knows the result of a given rally, nor indeed how he is in a better position to know the result than any of the other spectators, for his word goes. If intentions are fixed in a similar manner, then likewise we do not have appeal to special modes of cognition to account for privileged

self-knowledge (and indeed, there is nothing to cognise): the authority in question is a direct product of the ontological status of intention.

To establish the result more firmly, Wright gives a more detailed characterisation of the first-person epistemology of intention, and a more precise diagnosis of why judgement-dependence is the correct explanation for it. In order to facilitate the discussion Wright introduces what he calls the 'order-of-determination test'. To pass this test, a property must be objective; that is, the truth-value of the judgement that x is F must be independent of anyone's judgement that F is x. If a property passes the test, then since there is no logical connection between judgement and fact, it is a contingency that a given true judgement about F-ness is true. In such cases, judgements *track* the truth. In contrast, for a class of judgements to fail the order-of-determination test, it is not always a contingency that such judgements are true. Here judgements do not track the truth, but *determine* the truth.

To capture the idea of judgement-dependence, and of what it is to fail the order-of-determination test, more clearly, Wright considers the distinction between primary and secondary qualities. Secondary qualities are, traditionally, dispositions to produce certain experiences in us, so that to be red is to have the power to produce a certain qualia in a normal viewer under normal conditions. This idea may be captured with the following provisional (or provisoed) conditional:

$$C(\text{Jones}) \Rightarrow (\text{Jones would experience x as red} \Leftrightarrow \text{x is red})$$

where "C(Jones)" states that Jones operates under ideal conditions - in this case stipulating that Jones is a normal observer (has statistically normal perceptual equipment), in ideal viewing conditions (the conditions that typically obtain at noon on a cloudy summer's day).[10]

Wright is concerned to modify this basic model in such a way that is may be applied to properties which do not have a distinctive attendant phenomenology (whilst the basic model appeals to the qualia of experiencing something as red, there is no correlative 'experiencing

---

[10] Over the course of the cited papers, Wright gives several slightly different versions of the provisional conditional. One concern raised by these variations is whether the item that (helps) determine its own truth is a judgement, a disposition to judge, an avowal, or a disposition to avow (see for example 1989d , p. 632). A related question is whether the core conditional is a biconditional or not (in all cases the claim is that my judgement that Fx is sufficient to determine that Fx, but is it also necessary?). Such details have to be settled in a completed account, but since the whole project is somewhat programmatic it would be unwise to focus on the exact formulation of the conditional at the expense of the underlying principles.

something as an intention'). He therefore replaces the notion of a distinctive experience with that of judgement, with the justification that we can augment the C conditions in such a way to ensure that the subject will *judge* that x is red precisely when it is experienced as red. The result is the following conditional:

C(Jones) ⇒ (Jones would judge that x is red ⇒ x is red).

On the revised model, there is no more to being red than to be judged to be red by a competent agent under suitable conditions.

The adequacy of this type of account of colour is not our concern. What it provides is a means of capturing the idea that the extension of a predicate may be dependent on the way people apply it, that the concept in question is judgement-dependent.

To generalise, the form of the 'provisional conditional' is:[11]

C(Jones) ⇒ (Jones would judge that Fx ⇒ Fx).

If such a conditional is to capture a conceptual connection between judgement and fact, then the conditional must be *a priori*. However, not every *a priori* conditional of this form will support a claim about judgement-dependence, and so the account must be extended to include some additional criteria:[12]

(i) The conditional must be *a priori*

(ii) The conditional must be non-trivial.

(iii) The C conditions must be logically independent of the class of concepts under consideration.

(iv) There is no means of explaining the a priority of the conditional other than in

---

[11] In its crudest form the idea that judgements determine truth would be simply that whenever it is judged that P then P is true. It might be thought that this is untenable because it leads back to a collapse between what seems right and what is right - in which case no content can be given to the judgement in question. Whilst the inclusion of the 'proviso' (i.e. the C conditions) does ensure that no such collapse occurs, this in itself is not the motivating idea (it is certainly not Wright's reason). Rather, the aim is *to make sense* of the idea that under certain conditions what seems right is right in virtue of seeming right. If this basic idea is accepted, then whether it applies with or without qualification is not a pivotal issue.

[12] See Wright (1989c, p. 248).

terms of judgement-dependence.

Each condition is well motivated. For any set of judgements, an *a priori* conditional can be created by simply stipulating that the C-conditions are conditions under which Jones makes correct judgements. Hence, the stipulation of non-triviality is necessary if the *a priority* is to reflect a constitutive relation between judgement and fact. The third condition is required in order to ensure that it is Jones's judgement which determines the extension of F. For example, should the C conditions themselves entail that x is F, then the truth of the RHS of the conditional need have nothing to do with Jones's judgement, in which case we lose the idea that the extension of F is determined by Jones's opinion on the matter. Condition (iv) is needed to discount cases like that of pain, where the truth of Jones's judgement is guaranteed by the introspective availability of its subject matter.

What we are working up to is the idea that there is a provisional conditional of the form:

C $\Rightarrow$ (Jones judges that he intends to $\varphi \Rightarrow$ Jones intends to φ)

which meets Wright's criteria. The initial prospects are good, for the fact that avowals carry first-person authority makes it reasonable to accept such a conditional as *a priori*: after all, this authority consists in a general presumption that a subject's views about her own mind are right. And the fact that introspection is not the source of this authority rules out the most obvious alternative explanation. To complete the project, it has to be shown that there are non-trivialising C conditions, and that there is no alternative explanation for the a priority, quite apart from judgement-dependence.

Turning first to the existence of C conditions, it is significant that the authority of avowal is not absolute, and we can accept that self-appraisals can be mistaken. The aim of the C conditions is to stipulate conditions under which error cannot occur, and to this end Wright's strategy is to enumerate those circumstances which we think could lead to error, and then to state in the C conditions that such conditions do not obtain. Wright (tentatively) produces the following list of C conditions:

> [(1)] grasp of the appropriate concepts, [(2)] lack of any material self-deception or anything relevantly similar, and [(3)] appropriate attentiveness. (Wright 1989c, p. 251)

Conditions (1) and (3) are straightforward. If the subject is to form a judgement on the relevant matter he must be equipped with the concepts necessary to make that judgement, and we want his judgement to be the result of diligent consideration of the relevant issue.

Condition (2) is somewhat more complicated. Self-delusion - the irrational adoption of a belief one is in a position to know is false - though intrinsically puzzling, may arise with respect to any type of judgement, including judgements about one's own intentions. Also, as Wright alludes, there may be little correlation between the actions of the mad or the inebriated and their avowed intentions. Depending on the precise nature of the case, we may well conclude that the subject has intentions of which he is not aware, or that he does not have the intentions he claims he has. It is to allow for such situations that Wright is forced to include the 'no self-deception' criterion.

Unfortunately, the 'no self-deception' clause violates the requirement that the conditional overall be non-trivial; for stating that the subject is not self-deceived is simply another way of saying that the subject's judgement is right. There is no obvious means of excluding the possibility of such error in a non-trivialising manner, and so Wright accepts that C conditions which deliver an *a priori* true conditional cannot be given. However, Wright submits that the difficulty is not fatal for his project, for the provisional conditional can be modified in such a way that the 'no self-deception' condition is no longer required, and yet the conditional still supports a substantial conclusion. He says:

> The suggested measure proceeds on the observation that the absence of self-deception is "positive-presumptive". By that I mean that, such is the 'grammar' of ascriptions of intention, one is entitled to assume that a subject is *not* materially self-deceived, or unmotivatedly similarly afflicted, unless one possesses determinate evidence to the contrary. Positive-presumptiveness ensures that, in all circumstances in which one has no countervailing evidence, one is *a priori* justified in holding that the no-self-deception condition is satisfied, its trivial specification notwithstanding. (Wright 1989c, pp. 251-252)

To bring this point out we can split the conditions C into those which are trivialising (C_t) and those which are not ($C_{\neg t}$).[13] The provisional conditional is thus:

C_t & $C_{\neg t} \Rightarrow$ (Jones believes that he intends to $\varphi \Rightarrow$ Jones intends to φ)

---

[13] The notation here is borrowed from Sullivan (1994).

This conditional is of the form A & B $\Rightarrow$ C, which is equivalent to A $\Rightarrow$ (B $\Rightarrow$ C). It follows that the truth of A guarantees the truth of B $\Rightarrow$ C, and consequently that the degree of credibility given to B $\Rightarrow$ C must be at least as high as that given to A. Thus, if A & B $\Rightarrow$ C is true, then we can 'delete' the antecedent A to leave a conditional (i.e. B $\Rightarrow$ C) which is *as plausible* as A.

Applying this procedure to the provisional conditional, the trivialising C_t conditions can be deleted to give:

$C_{\neg t}$ $\Rightarrow$ (Jones believes that he intends to φ $\Rightarrow$ Jones intends to φ)

which must be *as credible* as the deleted C_t.

In claiming that the trivialising 'no self-deceit' condition is positive presumptive, Wright is claiming that we are *a priori* justified in believing that people are not liable to make mistakes when identifying their own intentions. If this is so, then the conditional above which results from the deletion of the trivialising condition will also be *a priori* credible.

Although we are no longer dealing with a provisional conditional which is *a priori* true, Wright finds the notion of *a priori* credibility equally useful as a means of establishing judgement-dependence. Certainly, the *a priori* credibility of the conditional cannot be explained away as a triviality, for the trivialising C-condition has been deleted. So the question is: why is the post-deletion conditional *a priori* credible? Why is it *a priori* reasonable to suppose that people can correctly identifying their own intentions? The standard answer - that the intention is an item in consciousness, and that it is a necessary truth that people know their own minds - is still not an option. In the absence of a better account, Wright observes that:

> The matter will be nicely explained if the concept of intention works in such a way that Jones's opinion, formed under the restricted set of C-conditions, play a *defeasible* extension-determining role, with defeat conditional on the emergence of evidence that one or more of the background, positive presumptive, conditions are not in fact met. (Wright 1989c, p. 252)

We can assume that one's own verdict is correct, unless there is firm evidence that it is wrong. But this possibility of error does not indicate that there is, after all, a cognitive process at work here. The situation can be explained by the fact that one's own judgement is extension-determining *unless* there is evidence showing that the verdict is wrong. By making

the extension-determining judgement subject to defeat, it can be explained how error is possible, but that nevertheless we have the right to assume an avowal is correct without the need for prior confirmation. Judgement-dependence would indeed explain the *a priori* reasonableness of the provisional conditional.

Sullivan's Objection

At least one author has found the argument so far to be entirely spurious, and a brief consideration of this view is useful to bring out the nature of Wright's argument. Sullivan (1994) holds that the *a priori* credibility of the provisional conditional is not something which stands in need of such a substantive explanation. As we have seen, the provisional conditional is the product of 'deleting' the positive presumptive antecedents from a trivially true (and hence trivially *a priori*) conditional.[14] In response, Sullivan says:

> Wright's conclusion is...that the *a priori* credibility of [the provisional conditional] is something that stands in need of explanation. And while we could accept that explanatory need, like *a priority*, is a feature transmitted through valid argument, this will not yet justify that claim, since [the pre-deletion conditional], as an acknowledged triviality, had no explanatory deficit to transmit. So where does the explanatory deficit come from? (Sullivan 1994, p. 158)

Sullivan's claim is that since the trivial conditional does not stand in need of explanation, any 'explanatory deficit' attached to the non-trivial conditional, produced by the act of 'deletion', is an illusion created by that very process. If the a priority of the trivial provisional conditional needs no explanation, then the *a priori* credibility of the (post-deletion) provisional conditional needs no explanation either. And if there is no need to explain the *a priori* credibility of the trivial provisional conditional, then there is no need to invoke the idea of judgement-dependence. As Sullivan says:

> When the various conditions under which an avowal of intention indicates an intention are "deleted" their influence is not thereby dismissed....It is only by deleting these conditions from our mind, as well as on paper, that we create the appearance of an explanatory deficiency. (Sullivan 1994, p. 158)

It turns out, though, that Sullivan's charge against Wright's argument cannot be sustained, for he misconstrues the significance of the deletion process. Certainly, Sullivan is right in saying that the a priority of this conditional:

---

[14] In this passage Sullivan refers to *a priori* credibility, not of the provisional conditional itself, but of the core conditional that if Jones avows that he intends to P, then he intends to P. That is, Sullivan carries out the process of "deletion" to its limit to remove all the C-conditions. As we shall see, Wright does treat *all* the C conditions as positive presumptive (1989c, p. 253), but deletion is only *necessary* (to avoid circularity) in the case of the trivialising 'no self-deceit' condition.

$C_{\neg t} \Rightarrow$ (Jones believes that he intends to $\varphi \Rightarrow$ Jones intends to φ)

is explained completely by noting that it is the product of a trivially true conditional following an act of 'deletion'. What Sullivan ignores is that in the explanation the a priority of the post-deletion conditional, essential reference is made to the notion of positive presumptiveness. The deleted conditional is the product of a triviality *plus* a claim of positive presumptiveness; and so to fully explain the *a priori* reasonableness of the conditional, it is necessary to explain *both* the process of deletion *and* the positive presumptiveness of the deleted condition. Certainly, the fact that the trivialising C condition is positive presumptive is something which stands in need of explanation (why can we assume an absence of self-deception?), and it is this which is the source of the 'explanatory deficit'. And it is precisely to explain why the absence of error can be assumed, without the need to first gather evidence, that Wright invokes the notion of judgement-dependence. Sullivan is, then, right that we should not forget the influence of any C condition just because it has been deleted, but by the same token we should not forget the process by which it was removed. It is only because he overlooks this that Sullivan finds the whole argument to be contrived.

The Temporal Spread of Judgement-Dependence

The argument up to this point has focused on judgements about the identity of one's own intentions. Our interest in the project lies with the eventual conclusion that what satisfies an intention is determined in an on-going fashion. To establish this conclusion, Wright first notes that it is not only judgements about the intentions one presently has which are extension determining, but also present judgement may be constitutive of past intention. He says:

> Reflect that the picture, if good at all, ought also to extend to knowledge of our former intentions or - what comes to the same thing, barring mistakes about other matters - knowledge of what currently would comply with them. For making sense of others' behaviour depends as much upon knowledge of their former as of their present intentional states; and, once again, our practice is to give avowals of former intentions, and judgements about what courses of action now comply with them, the same kind of default authority. (Wright 1989d, p. 633)

Hence, we get a corresponding trans-temporal provisional conditional:[15]

---

[15] Wright suggests that the C conditions must be suitably modified "so as to include at least the proviso that there is no major disorientation of memory at work." (1989b, p. 402). This appears to be another trivialising condition (i.e. Jones remembers correctly), and so must be positive presumptive, and hence a candidate for deletion, for the argument to work.

C ⇒ (Jones judges that he intend*ed* to φ ⇒ Jones intend*ed* to φ).

As Wright notes (for example 1987, p. 402), to identify an intention is to identify what fits it, and so we also have:

C ⇒ (Jones judges that φ accords with his prior intention ⇒ φ accords with Jones's prior intention).

If Wright is correct here, since what satisfies a former intention is determined retroactively by present judgements, then in the case of a prolific intention - an intention satisfied by a series of actions - this is sufficient to yield the intended result of on-going determination.[16]

Knowledge, Authority and Reliability

As mentioned above, this result is sufficient to undermine objective rule-following in its own right. The general structure of the argument is important for the additional reason that Wright suggests that it can be applied directly to rules. Before examining that step, though, I want first to show what is wrong with the initial thesis as restricted to intention.

The fatal defect in the theory is that it fails in its *raison d'être*, which is to account for the first-person knowledge of intention. The difficulty stems from the initial means Wright uses to bring out the fundamental problem of self-knowledge. He does this by contrasting two types of psychological state, each with a characteristic first-person epistemology. The first is that of an occurrent, phenomenological state, such as a pain, which is knowable by introspection, and where the identity of the state is not bound up with any disposition to behaviour. The other exemplar is a dispositional character trait, such as modesty or bravery, where the ascription of such properties - even to oneself - can only be made on the basis of manifest behaviour. With properties in this latter category, there is no first-person authority, and self-ascriptions are justified on the same evidence as is available to a third party.

---

[16] If both contemporaneous and retrospective judgements are extension determining then the possibility arises of a conflict between the two (say if I previously said that I intended to φ, and now say that I intended to ψ). Precisely how such conflicts are to be resolved - what order of priority exists between them - is something to be provided by a full account, but the mere possibility of conflict is not fatal to the overall theory.

Wright claims that intentions "straddle" these two paradigms. On the one hand intentions are knowable with first-person authority; avowals are quite properly made without reference to actual or dispositional behaviour. In this sense the epistemology of intentions resembles the epistemology of sensations. On the other hand, intentions are also 'dispositions-like', in that the ascription of an intention to someone is 'answerable' to subsequent behaviour. As we have seen, the difference between meaning *plus* and meaning *quus* is not something which is phenomenologically evident. Nevertheless, we should expect the difference to become apparent when the subject reaches the point in the number series where the functions diverge. *Ceteris paribus*, we can expect someone who means addition to accept additions, whereas someone meaning a quus-like function will give different answers. This is not to say that meaning *plus* or *quus* consists in a disposition, merely that the difference is only betrayed in the unfolding course of events. The fact that neither paradigm gives a satisfactory account of the epistemology of intention obliges us to formulate a third option, which is the cue for judgement-dependence.

In saying the epistemology of intention 'straddles' the two paradigms, Wright highlights elements of the respective routes to knowledge. When forming verdicts about one's own intentions, the process is like that with phenomenological states, in that it involves no external observation; but in common with dispositional states, introspection cannot be the belief-formation process involved.

There is, though, another sense in which there is a 'straddling', this time concerning the *status* of self-knowledge. As with phenomenological states, it is *a priori* that I know the contents of my own mind, for it is my mind. Yet reference to the answerability of such judgements brings with it an *a posteriori* element. As Wright says:

> It is part of regarding human beings as persons, rational and reflective agents, that we are prepared to ascribe intentional states to them, to try to explain and anticipate their behaviour in terms of the concepts of desire, belief, decision and intention. And it is a fundamental anthropological fact about us that our initiation into the language in which these concepts feature results in the capacity to be moved, who knows exactly how, to self-ascribe states of the relevant sorts - and to do so in ways which not merely tend to accord with the appraisals which others, similarly trained, can make of what we do but which provide in general a far richer and more satisfying framework for the interpretation and anticipation of behaviour than any which they could arrive at if such self-ascriptions were discounted. (Wright 1989b, p. 402)

By "richer and more satisfying framework for the interpretation and anticipation of behaviour", I do not see what Wright can mean other than "pragmatically superior". That is,

taking into account other people's self-ascriptions often provides a more accurate means of predicting their behaviour than is available if avowals are not taken into consideration.

There is a danger in that it might be thought, because this "fundamental anthropological fact" is *a posteriori*, that consequently the authority of avowal is similarly *a posteriori*. In that case, knowledge of one's own intentions would be deemed to be both *a priori* and *a posteriori*, which suggests that something must already have gone wrong.

It would, though, be a mistake to dismiss Wright's argument so readily. The reliability here is without question an *a posteriori* matter, and in the abstract, reliability with respect to a class of judgements is respectable grounds on which to say that the subject knows what he is talking about. The fact is that reliability may be sufficient to confer authority on inductive, empirical grounds, but it does not follow that the authority avowals have actually *relies* on such *a posteriori* grounds. For example, it is an empirically ascertainable fact that a tennis umpire is reliable in his judgements (i.e. if we compare his stated verdicts with the records for the match), but it is not *because* he is reliable that he has authority; it is the other way around.

Actually, the only problem is our insistence in seeing things as either *a priori* or *a posteriori*, and in this respect Wright's thesis is entirely successful in providing a 'straddling' alternative. Under the thesis of judgement-dependence, self-appraisals can be assumed to be right *a priori*, but such *a priori* claims of knowledge are defeasible on the basis of subsequent behaviour - that is on empirical grounds. Is such self-knowledge *a posteriori* or *a priori*? There is, of course, no clear answer: the solution has elements of each, and so does indeed straddle the two positions.

Nevertheless, the mere reconciliation of the tension between the *a priori* and the *a posteriori* is not sufficient for Wright's aims: he has to do more if he is to account for our knowledge of our own intentions. As noted, it is an *a posteriori* fact that avowal is a useful guide to subsequent behaviour. If it were not, then we should repudiate the knowledge claim. That is to say, reliability as measured against subsequent behaviour is necessary if claims of self-knowledge are to be sustained. As Wright says:

> What determines the distribution of truth-values among ascriptions of intention to a subject who has the conceptual resources to understand these ascriptions and is attentive to them are, in the first instance, nothing but the details of the subject's self-conception in relevant respects. If the assignment of the truth-values, so effected, generates behavioural singularities - the subject's

> behaviour clashes with ingredients in his/her self-conception, or seems to call for the inclusion of ingredients which he/she is unwilling to include - then the self-deception proviso...may be invoked, and the subject's own opinion, or lack of it, overridden. (Wright 1989c, p. 253)

Hence it is only when the subject says he intends one thing, and subsequently does another, that his own verdict may prove to be incorrect.

The message that emerges is that the epistemology of intention must not only account for *a priori* authority, it must also account for the *a posteriori* reliability of avowal - reliability, that is, when measured against subsequent behaviour. Since the reliability of one's judgements depends entirely upon the frequency with which such 'clashes' occur, to explain self-knowledge it is necessary to explain the general absence of 'clashes'.

Wright acknowledges that an explanation of reliability is indeed a desideratum of his thesis:

> What needs explanation is rather how a subject can be so much as in a position to be *reliable*? If proof of the pudding is in subsequent performance, what basis can there be for an opinion at a stage at which a third party can have none? (Wright 1989b, p. 400)[17]

And, as it transpires, Wright is happy to credit his theory with this very accomplishment:

> Explaining the *a priori* reliability of a subject's C conditioned beliefs about his intentions will do nothing to explain the reliability of his avowals - even assuming our right to assume they are honest - unless the C conditions in question are likely to be met. But there seems to be no cause to anticipate problems on that score. Attentiveness - however precisely it should be elaborated - is presumably, like lack of self-deception, a positive-presumptive condition; and a subject's possession of the appropriate concepts is a prerequisite for their being able to effect the avowal in the first place. So there is every promise of a straightforward kind of explanation of the authority which avowals of intention, *qua* avowals, typically carry. (Wright 1989c, p. 253)

If Jones goes around making avowals in situations when the C conditions are not met, then the provisional conditional will not be extension-determining, and Jones would not then be reliable in his judgements. However, as Wright notes (1989c, p. 253) there is no difficulty in the idea that the remaining C conditions - appropriate attentiveness and concept possession - are frequently satisfied.

It is telling that in the above quote Wright starts by considering the explanation of *reliability* - and an *a priori* reliability at that - but he concludes that there are no serious difficulties

---

[17] This passage refers directly to avowals not involving intentions, which Wright portrays as not having special first-person authority. Nevertheless, from the context there is no doubt that the question of reliability is intended to apply to avowals of intention as well.

facing the explanation of *authority*. This would be unobjectionable if our conditional were *a priori* true: with a true conditional, satisfaction of the C conditions guarantees that Jones's verdict is right, and frequent satisfaction of the C conditions would entail that Jones's judgements are frequently right. But we do not have a true conditional. All we do have is the right to accept Jones's judgement as right *in the absence* of countervailing evidence. The very reason for the shift from truth to credibility was to allow for error, with the rider that the absence of error can always be assumed, unless we have positive evidence to the contrary.

Unfortunately, whereas Wright claims to explain the *a priori* reliability of avowal, all he actually explains is why, on a case by case basis, it is *a priori* reasonable to *assume* the avowal is right. Nothing has been said which has any bearing on the frequency with which clashes do, or do not, actually occur, which is to say that the explanation of reliability is missing.

The point can be illustrated using Wright's own example of the tennis umpire. In tennis, whether a ball is in or out is subject to the decision of the umpire. If the umpire were the final arbiter, then we should have a complete explanation of why her judgements are reliable, this explanation simply being that whatever the umpire says goes. Yet, as Wright notes, the umpire's verdict can be overruled by a higher official, and so the umpire's verdicts carry only a defeasible authority: what the umpire says goes *unless* the match referee intervenes. The possibility of defeat makes a vital difference when it comes to explaining the reliability of the umpire's verdicts, for it is not then sufficient to note the default authority which her decisions have. If the match referee habitually overrules the umpire, then although the umpire still enjoys a default authority, she would not thereby be a reliable guide to the way in which points are actually awarded to the players. To explain the umpire's reliability (when it occurs) it is necessary to give not only an account of the hierarchy of authority involved, but also to explain why the match referee tends not to overrule the umpire.

In more general terms, the point is that an explanation of a subject's default *authority* is not sufficient to explain his *reliability*. The difference between the two things is clear. To have authority means that, as a default, it can be assumed that you are reliable. To actually be reliable, though, requires the frequent absence of countervailing circumstances. Explaining why we have the right to assume someone knows, and explaining how they do actually know may, therefore, be very different things. In merely describing the defeasible authority of avowal, and the nature of the defeat conditions, Wright does not explain why these defeat

conditions - 'clashes' - are uncommon, and so his theory fails to explains the reliability of avowals, and, consequently, the possibility of self-knowledge.[18]

The difficulty which Wright's theory faces in the specific context in which it is applied is really a re-surfacing of the initial fundamental problem of self-knowledge: how is it that our avowals and our behaviour are in harmony, respectively betraying (for the most part) the same intentions? To explain this it is imperative to identify a relevant connection between self-appraisal and subsequent behaviour. (In the Cartesian case, the connection is one of common cause.) The reason why judgement-dependence cannot explain reliability is simply that it fails to incorporate any such relation. If intentions are, broadly speaking, dispositional, then, just as they cannot be non-dispositional, occurrent objects of inner contemplation, they cannot be equally non-dispositional judgement-dependent states. (Dispositions do not arise from the mere act of judging.) Judgement-dependence, in precluding an intention from being the kind of property which can enter into the causal nexus, ensures that any correlation between the two is merely accidental. Whatever the correct account of self-knowledge of intention is, it is not the one which Wright gives us.

Judgement-Dependence and Rules

It remains to be seen whether the argument for judgement-dependence can be transferred from intentions to rules and meaning. Despite the failure of the argument when applied to intentions, this project is still worth pursuing, not least because upon investigation it transpires there is a significant difference between intentions and rules/meanings which makes the case for the judgement-dependence of the latter that much stronger.

As Wright says, the thesis about meaning:

> does not require construal of meaning as a kind of intention; it is enough that the concepts are relevantly similar - that both sustain authoritative first-person avowals, and that this circumstance is to be explained in terms of the failure of the order-of-determination test. (Wright 1989c, p. 254)[19]

---

[18] Interestingly, the flaw is not inherent within the overall model. Continuing with the example of the tennis officials, the reliability of the umpire is readily explained by the fact that both the umpire and the official attempt to respond to the location of the ball, have sound perceptual faculties, and are in positions offering superior views of the court. Given these facts, we have every reason to expect a general correlation between their verdicts. The reliability here is explained with reference to judgement-dependence, although not exclusively in such terms.

[19] In addition see Wright (1989c p.257). Puzzlingly, though, Wright also states that when grasping the meaning of a word 'in a flash', or coming to understanding how to continue a series "there is no institution of avowal in the strict sense - the subject's word carries no *special* authority." (Wright 1987, p.400).

Wright does not fully develop this project,[20] but the framework is in place, and it is readily seen how the argument ought to proceed. First we should formulate a provisional conditional for rules, admitting that Jones has a certain authority as to which rule he follows. It would then be necessary to argue that this authority cannot be the result of an act of introspection, and that the only explanation for it is in terms of judgement-dependence. There is now an additional requirement, for it should also be ensured that the objection raised above to the judgement-dependence of intention (concerning the explanation of reliability) does not resurface to discount the thesis about rules as well. If, further, the subject has retrospective authority about the application of his rule, and this too is to be explained in terms of judgement-dependence, then on-going determination would be established.

This project faces an immediate difficulty. The aim is to see whether a conditional like this:

C $\Rightarrow$ (Jones judges he grasps the addition rule $\Rightarrow$ Jones grasps the addition rule)

is *a priori* true (or *a priori* credible), and if so whether this a priority is best explained in terms of judgement-dependence. Yet to form the belief that he grasps the rule for addition, Jones must have the concept under investigation (namely *addition*), and hence *already* grasp the rule in question, for grasp of the rule is constitutive of having the concept. It cannot then be *because* he makes the judgement that Jones has the rule; possession of the rule is a *prerequisite* for the ability to make the judgement. As a result, the model cannot be applied to judgements about the overall identity of the rules one follows.

This, though, does not signal the end of the investigation, for consideration of judgements about the overall identity of my rules does not exhaust the scope for the application of Wright's thesis. In particular, the very judgement which is targeted by the indexical argument - that I am following the same rule as before - ought to be examined in this light.[21] The relevant conditional is this:

---

[20] Wright says: "I would like to be in a position to offer a supported opinion about whether this can be approached along the lines sketched for the cases of colour and intention. But I have, at the time of writing, no settled opinion to offer about that, let alone about whether the idea can ultimately be made good." (Wright 1989c, p. 257)

[21] Wright explicitly identifies this as a credible application of the model. He says: "Challenged to justify the claim that I formerly meant addition by 'plus', it will not be necessary to locate some meaning-constitutive fact in my former behaviour or mental life. A sufficient answer need only advert to my present opinion, that addition is what

C ⇒ (Jones judges he is following the same rule as before ⇒ Jones is following the same rule as before).

Here, unless the target rule requires us to make the judgement "I am following the same rule as before", the circularity mentioned above does not re-occur. Whilst this does leave something of a singularity - it is not clear how the thesis of judgement-dependence could be applied to judgements involving the concept of rule-following itself - in most cases the problem does not arise. Since judgement-dependence is not immediately precluded in the majority of cases, the prudent course of action is, I suggest, to accept that whilst there are difficulties which a complete theory will have to address, the initial proposal remains worthy of further consideration.[22]

There is an additional, related concern, that the mere mention of concepts within the C conditions makes the account circular. That is, whilst there need be no *particular* rule which is both mentioned in the C conditions and required for the making of the relevant judgement, the model presupposes that *some* judgements can have content quite independently of judgement-determination. We have to assume that there is some judgement which has a determinate content in order to get the account up and running.[23]

Again, though, this worry does not warrant the outright dismissal of the project, for as long as the *specific* content which is being analysed in terms of judgement-dependence is not presupposed by the C conditions, then the account is not circular. The argument can be put like this. Let us assume that some specific concepts can be ascribed to Jones, without comment as to their constitutive nature. If, under this assumption, if follows that we have a conditional which satisfies all of Wright's requirements, then we should have shown that

---

I formerly meant, and still mean, and to the *a priori* reasonableness of the supposition, failing evidence to the contrary, that this opinion is best." (1989c, p. 254).

[22] The only other judgement which might have some bearing on the issue is as to the requirement of my present rule. There is a strong intuition that if Jones holds that 1004 follows 1002 according to the rule he has been following, then, if perceptual errors and the like can be excluded, his word is to be accepted on the matter. It is reasonable to accord such judgements some degree of first-person authority, but in this case there is no bar to an explanation in terms of introspection. For with a particular judgement, involving a single application of the rule, there is nothing to discount an epistemology based on introspection. For all the considerations raised, it could be that when one follows a rule, one identifies the requirements as and when they are needed. Judgements about application do not answer to subsequent behaviour in any way, and so there is no objection to the idea that the requirements of the rule can be brought into consciousness as and when desired.

[23] This point is raised by Boghossian (1989b, p. 547).

judgements *using* these concepts are extension-determining. In the first instance, the fact that we assume that some concepts are objectively determined does not interfere with the conclusion that others are not. In addition, the fact that we assume that Jones is in possession of the concepts required to make certain judgements is not to assume that these concepts are objective. As already noted, it is a problem to see how judgement-dependence can be applied across the board, but the difficulty is not substantial enough to end the project before it gets going.

Turning, then, to questions of detail, we need to know whether the conditional meets Wright's criteria. We have just established that there is no obvious reason to suppose that C conditions cannot be given without circularity. It remains to be seen whether there are C conditions, specified in a non-trivialising manner, which give an *a priori* true/warranted conditional; and if so, whether the only explanation for this come in terms of judgement-dependence.

First off, there is a *prima facie* case for first-person authority with respect to the relevant judgements. Consider someone who continues the series 1000, 1002, 1004, 1008, 1010. On being notified that he has make a mistake (1006 should follow 1004), the agent becomes quite adamant: his initial answers were correct, that the rule *he* is following requires 1008 and not 1006. In this situation, we should have no option but to accept the subject's view on the matter. If he thinks that 1008 accords with the rule he has been following all along, and we cannot explain this abnormal response as a mistake due to perceptual errors or the like, then we have no basis on which to disagree. Although it looked as though he were following the rule add 2, there is no reason why he should not have been following a different rule all along, with the difference becoming apparent only at the point now reached. If the subject says he is following the same rule he has always been following, then even though his behaviour is not what we expected, we should accept his verdict. In other words, it is at least *a priori* reasonable to take the relevant avowal at face value, making it highly plausible that there is a provisional conditional of the type required which is itself *a priori* reasonable.

We should not, though, expect any such conditional be *a priori* true. Major disruptions in memory are possible, and as with any psychological state, so too is self-delusion. To ensure a correct verdict, both situations must be discounted by the antecedent C conditions, yet both are trivialising: saying you have not forgotten is another way of saying that you are right. Nevertheless, it can be accepted that the absence of both such interfering factors can be

assumed unless there is definite evidence to the contrary. In other words, it is positive presumptive that neither situation occurs, so that Wright's technique of deletion can be utilised here too. Upon deletion of these conditions, we are again left with a non-trivial *a priori* reasonable conditional, with the remaining C conditions being as before, that the subject has the necessary concepts, and is appropriately attentive.

How, then, is the *a priori* reasonableness of the conditional to be explained? Certainly an explanation in terms of introspection is not an option. To introspect that I am following the same rule as before, I should have to identify both past and present rules to make a creditable comparison between them. But as ever, the limitless applications of a rule cannot all be scrutinised, and so neither the identity of the present rule, not the identity of the past rule, can be ascertained on this basis.[24]

As a result, we are faced with a reasonable candidate for judgement-dependence: an *a priori* warranted conditional of the correct form, the a priority of which cannot be explained in terms of introspection. The way is open for an explanation in terms of judgement-dependence. First, though, we should ensure that same problem which arose for intentions - that of accounting for reliability - does not re-occur in the present context.

It is in this respect that there is a marked difference between the two cases. The point comes out when we consider the situation in which one's avowal has to be rejected. With intention, were I to say that I intend to φ, but in fact I subsequently ψ, then the manifest 'clash' between avowal and behaviour overturns my avowal. This type of situation proved troublesome for Wright's thesis because a knowledge claim requires a strong correlation between avowal and behaviour, one which cannot be expected to occur by chance. A satisfactory epistemology of intention had to explain the frequency with which such correlation occurs, and it is due to its failure to provide any such explanation that Wright's thesis was rejected.

Turning to rules, the possibility of a 'clash' between behaviour and avowal is also ever-present. Bearing in mind that the judgement in question does not specify the identity of the rule I am following, if I am to follow the same rule, I must at least presently be following

---

[24] The fact that one's prior rule is in the past and hence not available for current introspective scrutiny does not in itself preclude an explanation in terms of introspection, on condition that the prior rule could have been identified introspectively, and its identity remembered.

some rule. Yet, if I am at a complete loss to say what the rule requires in any given situation, then the only possible conclusion is that I am not in fact follow a rule at all. In this case my claim to be following a rule, and with it my claim to be following the same rule, is seen to be false. The mere claim that I am following some rule, without specifying which one, remains 'answerable' to subsequent behaviour: I must do *something* in the name of the rule if my avowal is to stand.

Even so, the central strength that the case for judgement-dependence gains in its application to rules (as opposed to intentions) is that with rules the type of correlation here required is relatively loose. This is due to the point noted above that that we are not now concerned with judgements about the specific identity of a rule, together with the familiar fact that any series of actions accord with some rule. It follows that so long as I come up with *some* answer as to how the rule I have been following ought to be applied, there is no 'clash' between my behaviour and my claim that I am following a rule, and my own verdict cannot be overridden.

This is not to suggest that it is *a priori* that genuine 'clashes' are rare. The point is rather that in light of the *vast* range of dispositions which do not override one's avowal - namely anything which looks like rule-following - then an explanation of our general reliability is not overly troublesome. It might just be that when we think we are following a rule we are disposed to give some kind of answer rather then none at all. Achieving consonance between avowal and behaviour is effortless, and so the explanatory agenda which arose for intentions does not arise for rules.

An Alternative Explanation

Even though we have an *a priori* reasonable conditional, one for which the problem of reliability does not arise, the argument for judgement-dependence still fails. Wright's final condition is that there should be no alternative explanation for the acknowledged situation, but in this case the *a priori* reasonableness of the provisional conditional has a quite mundane explanation.

What is to be explained is this: when Jones says "I am following the same rule as before", why should we accept his verdict at face value? On what basis do we disregard the possibility that he has lost his previous rule, and is now following another; or that he is now no longer following any rule at all?

To begin to formulate an answer, we can look again at the example used to motivate the claim that "I am following the same rule as before" has first-person authority. The envisaged situation was one in which someone follows a rule as we do, but then deviates from the expected course. The salient point is that initially - *before* the deviation - we are left in no doubt that a rule is being followed, and the evidence for this is nothing to do with the subject's avowal, but is manifest only in his behaviour.

Precisely what type of behaviour warrants the claim that someone is following a rule is not important for the moment.[25] On the assumption that there is some such evidence, it is notable that any reason we have to believe that someone is following a rule must support the claim that a single rule is being retained and followed over a period of time. Since a rule-follower will typically grasp more than one rule, this requires that the agent be able to re-select that specific rule - he must follow that rule and not some other - from his repertoire of rules. The evidence in question is, therefore, evidence that (a) a rule is grasped, (b) it is retained over a period of time, and (c) can be reliably re-selected at will as the rule which is in force.

If we have evidence (which does not involve avowal) that Jones can retain and re-select the same rule, then we have a basis on which to say that if Jones thinks he is following the same rule, then indeed he is. For to be a reliable rule re-selector means that when you intend to follow the same rule as before you can, and hence when you think you are following the same rule as before, in the main you are. So the situation described, where the ability to follow a rule in the past is not in question, we have a situation in which we have evidence which justifies the provisional conditional. Reason to believe the subject is a reliable rule re-selector is reason to believe that when he thinks he is following the same rule, he probably is.

The situation is then that we accept that Jones is a reliable re-selector of rules *given evidence* that he was a reliable re-selector of rules. There is little mystery about the a priority here. Although our evidence is itself *a posteriori*, *relative* to this evidence it is *a priori* (indeed, a triviality) that we should accept Jones as a reliable rule re-selector, and so there is no call for an explanation in terms of judgement-dependence.[26]

---

[25] An examination of the type of behaviour that warrants the ascription of a rule is conducted in Chapter 7.

[26] There is a possible objection to this criticism of Wright's argument which may be made at this point. The alternative 'mundane' explanation offered above relies on the inductive claim that, since the agent has demonstrated the ability to follow rules in the past, we can claim that she can continue to do so in the present.

This explanation covers only one type of case, in which evidence of the ability to follow a rule is central to the example. What happens when such evidence is not available? Do we still accept default authority? Certainly it is no accident that the example used above, where an acknowledged rule-follower deviates from the expected course, is required to motivate the thesis of first-person authority. For when we move away from this type of scenario, things become far less clear-cut.

The situation to be envisaged is one where Jones says "I am following the same rule as before", yet we have no concrete reason to believe that he was actually following a rule at the earlier time (suppose we just don't have any evidence either way). Should we then accept the avowal? One way of reaching a decision would be to note that the authority of "I am following the same rule as before" can be no greater than the authority of "I was following a rule", for if the latter is false, so too is the former. Further, the authority of "I was following a rule" can be no greater than the authority of "I am now following a rule", for we should not expect the passage of time to increase one's ability to identify aspects of one's past psychology. Hence, at the very least, "I am following the same rule as before" (in the absence of behavioural evidence) can only enjoy first-person authority if "I am now following a rule" carries default authority. So the pertinent question is: is this type of judgement, about my present activities, to be taken at face value without the need for corroborating evidence?

In fact, the absence of concrete evidence of rule-following alters the situation minimally, for in accepting the subject as capable as making an avowal we accept that he grasps, retains and

---

Thus, the mundane explanation depends on the fact that the agent exhibits behaviour which warrants the ascription of rule-following. However, since the indexical argument shows that objective rule-following is impossible, evidence that the agent is a rule-follower can at only be evidence that she is a 'creative' rule-follower. Our explanation of the apriority of the provisional conditional therefore involves the claim that such-and-such behaviour warrants the ascription of creative rule-following. This means that the mundane explanation is itself couched in terms of judgement-dependence, and Wright's fourth condition is satisfied after all.

What this objection neglects to consider is the point made earlier about where the onus of proof lies when it comes to making a claim about the subjectivity of rules as a response to the indexical argument. As discussed above, we cannot adopt a subjectivity thesis on the grounds that it is the only notion of rule-following which survives the indexical argument. To do so would be to appeal to subjectivity simply as a means of 'saving the phenomenon', a strategy which is unacceptably *ad hoc*. We have to have some reason other than that objective rule-following is impossible to suppose that rules are subjective. It is for this reason that we have been looking at an argument which claims to show that rules are judgement-dependent independently of the indexical argument. By the same token, we cannot use the conclusion that rules are not objective when defending Wright's argument, for otherwise the argument would no longer be independent of the indexical argument. This being so, the claims made in the above paragraph which arise from the indexical argument, namely that since rules are not objective they are subjective, and that consequently evidence of rule-following must be evidence of creative rule-following, are not ones which can be used in the defence of Wright's thesis.

reliably re-selects a number of different rules (namely those governing the words he uses). Whilst the fact that he can retain certain select rules does not entail that he can do the same across the board (there is no reason why someone should not be selectively amnesiac, and lose hold of certain rules, or certain type of rules, and not others), nevertheless, to accept an avowal we accept the subject is a competent linguist, and thus has the ability to follow a substantial number of rules. In the absence of evidence against, we should have the right to extrapolate this ability. On inductive grounds we should be warranted in saying that Jones is a reliable rule-follower, and this warrants acceptance of the claim that he is following a rule, and that he is following the same rule. The assumption that the subject is capable of making the avowal is then an assumption on which we can describe the subject as someone who can retain and follow the same rule at will, in which case we have good reason to accept that when he says he is following the same rule, he is. Again, the explanation is mundane, entails no ontological revision of rules, and on this basis alone displaces all talk of judgement-dependence.[27]

Conclusion

With the failure of judgement-dependence, we are left with no good reason to suppose that the continuing identity of a rule is fixed by one's own opinion on the matter. In the absence

---

[27] The rebuttal offered here is also effective against the earlier attempt by Wright (1986) to establish on-going determination on slightly different grounds. Wright suggests that if a community of speakers have all exhibited their competence in applying concepts which concern directly perceptible states of affairs (colours, shapes, sounds etc.) then, in the absence of any perturbing factors (funny lighting, trick mirrors etc.), we are necessarily justified in thinking that the (non-collusive) communal verdict over a series of judgements are correct. Thus if a group of people who have all shown good grasp of the concept red in the past are faced with a ball placed in an open field in good sunlight and, over a series of observations, all call the ball 'red', then we necessarily have reason to believe that the ball is red.

Wright notes that we can only have such a warrant if we can, on *a priori* grounds alone, discount the possibility that the members of the community should *lose* the ability to apply the concept correctly, and yet still non-collusively concord in their use of the predicate in any given case. That is, we can we discount the possibility that the ball in the example is in fact blue, and that our observers have all just failed to apply the word 'red' correctly by calling the blue ball 'red'. Wright suggests that the fact that this situation can be discounted shows that protracted, non-collusive communal agreement which is mistaken (in the absence of perturbing factors) is not logically possible. And the only way he can see to explain such a logical impossibility is if the on-going verdict of the community determines how the predicate ought to be applied as it goes along.

We can agree that the possibility that a number of observers should all just 'drift away' in unison from the requirements of a rule is not one that we should sanction. Hence, if a group of people who have competently identified red things in the past all agree over a period of time that a ball is red when there is no reason to believe that viewing conditions are abnormal, then I think we are justified in believing that the ball is red. However, a preferable explanation is that if people have displayed competence at applying a concept in the past, and if we have no reason to suppose that features in the environment interfere with this ability, then we are justified in thinking that they will *continue* to have this ability in the future, and to exercise it correctly. Therefore the warrant to believe that the communal verdict is correct is just the product of induction on the basis that each member in the community displayed competence at applying the concept in the past (a condition which Wright explicitly requires). There is no indication here of a conceptual connection between communal verdict and truth.

of alternative motivating arguments, we cannot rely on subjective on-going determination to contain the indexical argument. The Wittgensteinian creativity thesis cannot be sustained.

There is, though, a useful result to be secured from the foregoing discussion. The instigation for Wright's thesis is that there is a fundamental difficulty with the conventional epistemology of rule-following: on the one hand grasp of a rule can only be ascribed on the basis of manifest behaviour; on the other, the claim "I am following a rule" is not made - nor is it expected to be made - on the observation of one's own actions. For Wright the problem is to accommodate both characteristics. However, an accommodating solution is not the only option. A more direct approach would be to acknowledge that these two paradigms are irreconcilable, so that one of them must be rejected. In particular, it might be that one of the views about the epistemology of rules depends upon a defective, or more likely simply underdeveloped, view of the nature of rules. Once that situation has been rectified, we should not hesitate to alter the epistemological picture accordingly. In pursuing this strategy, the ensuing choice is not difficult to make. For it is only upon the systematic examination of rules that we realise that introspection is not applicable, and that it is possible to think that one grasps a rule and yet not grasp a rule (someone may think that they grasp a rule, but on execution of it, find themselves at a total loss as to how the rule ought to be applied). In light of these revelations (by Wittgenstein), it becomes quite reasonable to attribute any adherence we may have to first-person authority as the product of our failure to make these observations. Once the results are exposed, we should not hesitate to reject the first-person authority of avowal with respect to rules.[28]

Given that I cannot directly perceive (introspect) my grasp of a rule, and given also that my belief that I am following a rule does nothing to determine that I am indeed following a rule, there is little option but to conclude that I do not know directly that I do indeed grasp a rule. If I am to be justified in making such a claim, then there must be specific evidence on which it is made, and in turn, since it is not directly observational, the claim must be made on the basis of *inference*. As we shall see, this result, and the ensuing details of the nature of the inferential procedure involved, opens the way to an alternative response to the indexical argument, a response which is developed in the following two chapters.

---

[28] One way of putting this would be to describe Wright's argument as, broadly speaking, transcendental: taking the premise that we have a certain kind of self-knowledge (with respect to our own rules) he argues that such knowledge is possible only if rules are subjective - this subjectivity is thus a necessary condition for the existence of the knowledge in question. The response is correspondingly sceptical: the failure of the introspection model shows that such knowledge is not to be had.

PART THREE

MEANING WITHOUT RULES

6. Eliminating Rules

Taken at face value, the indexical argument shows that rule-following is impossible. Correspondingly, the most direct response to the argument would be to accept this conclusion as is, and to simply eliminate rule-following from our ontology. The main objection to this 'rule-elimination' is the thought that rules are essential for meaning. On the basis of the constitutive claim:

(CC) To grasp a meaning it is necessary to grasp a rule,

in eliminating rules, we eliminate meaning as well. But we cannot ever conclude on the basis of an articulated argument that meaning is impossible, for the very formulation and presentation of any argument depends on the possibility of language. It is the prospect of meaning nihilism which renders any argument against rule-following paradoxical, and which motivates the search for a position, such as irrealism or on-going determination, which *salvages* some notion of rule-following, and which thereby lessens the destructive effect of the indexical argument. As we have seen, irrealism is untenable, and on-going determination is unmotivated.[1] In the absence of further 'salvaging' options, I suggest that we relinquish the aim of saving rules, and instead accept the obvious conclusion: nothing determines which

---

[1] In Chapter 5, Wright's creative notion of rule-following was found to be an unmotivated response to the indexical argument. Given, though, that that position has not been shown to be incoherent, it should perhaps remain as an option. Indeed, intuitively the elimination of rules has to be the least favoured option, one which should only be endorsed once every alterative has been exhausted. Is Wright's thesis, then, not preferable to elimination simply because it is not eliminative?

The matter (still) comes down to one of motivation. In the case of rules, certainly every intuition is to reject elimination as wholly unworkable. It is for this reason that elimination seems to be the least favourable option. Without rules, content is impossible, so we must salvage rules somehow. If, however, it can be shown that elimination is coherent – as is attempted here – *the need to reconstruct rules is removed.* In addition, it should be noted that the damage done to the notion of a rule by the sceptical argument is enormous. Whilst it may be substantive to think of a rule being extended by new applications at the limits of its reach (the type of picture offered by strict finitism, for example), Kripke's insight (fully exploited by the indexical argument) is that at every moment, with every application, and with every re-application of a rule, nothing is fixed. The picture we have is not one in which we lay down a track that then remains fixed for others to follow; instead every time I come to follow rule, every time I start the series 2, 4, 6..., any answer is possible, no matter how far or how well I or anyone else followed the rule last time. If rules never dictate anyone's behaviour, if they never fix any response, what is the point of saving them? If we fully take on board the message of the indexical argument, the revisionary project is wholly unmotivated, and so unless there is some positive reason to suppose that created rules are essential for content, the *prima facie* conclusion that rule-following is impossible should stand. If viable, and other things being equal, elimination wins.

rule I grasp, therefore I do not grasp, or follow, a rule at all. In that case, the only means of avoiding meaning nihilism is to reject CC. For if rules are not constitutive of meaning, then eliminating rules does nothing to threaten the possibility of language. Rule-elimination, together with the rejection of CC, is, I propose, the correct conclusion to draw from the forgoing rule-following considerations. The purpose of this chapter is to establish the viability of this thesis.

Motivating the Rejection of CC

This method of avoiding meaning nihilism can only succeed on the understanding that the indexical argument is directed against the possibility of grasping a rule - a correctness condition - and not against the grasp of meaning *per se*. (If the indexical argument showed that meanings are underdetermined quite apart the supposed constitutive role of correctness, then eliminating rules would, of course, do nothing to save meaning.) That this is indeed the case is perhaps concealed somewhat by the initial presentation of the argument; for the claim was made that there is no fact of the matter as to whether I *mean green* or *grut*, and so the argument appears to affect meanings as much as it affect rules. However, the argument actually only shows that nothing determines trans-contextual identity for *correctness conditions*, and so only has bearing on meaning *on the assumption* that meanings determine correctness conditions. If this is assumption is not true - that is, if CC is rejected - then the indexical argument does nothing to threaten the possibility of language.

The indexical argument is, though, insufficient in itself to warrant rejection of CC. As it stands, we have a choice: either endorse CC, and end up with the incoherent conclusion that meaning is impossible, or reject CC. The former is position is untenable, which in itself seems to recommend that CC be rejected. However, the very motivation for the rule-following consideration is the established position that rules are necessary for content, which is to say that the received wisdom is that rejection of CC is itself incoherent. Whilst we do have a reason to *consider* the possibility that CC is false - it being the only potential means of avoiding incoherence - we cannot claim that rejection of CC is coherent *by default*. In short, we cannot reject CC in order to avoid the incoherence of meaning nihilism unless we can be sure that the rejection of CC is itself a coherent option.

Ideally, the way to proceed would be to show directly that the rejection of CC is a coherent position, perhaps with the construction of a counterexample to CC, or by giving a characterisation of meaning acceptable to all parties in which rules play no essential part. Yet

we should not expect any such counterexample, or characterisation, to be forthcoming. For if we could provide either, then we should be able to establish that rules and meaning are independent of each other *without* recourse to the rule-following considerations, in which case the whole investigation of rules would be redundant. It is precisely because there is such persistent conviction that rules are essential to meaning that the examination of rules, along with arguments which threaten the very possibility of rule-following, gain significance. If it were not so difficult to overthrow the accepted view that meaning is rule governed, then we should not bother with the rule-following considerations in the first place: the fact that the rule-following considerations are compelling indicates that convincing counterexamples or characterisations of the type sought are hard to come by.

An alternative strategy is therefore required. Rather than show directly that the rejection of CC is coherent, I propose to show that we have *no reason* to believe that CC is true. This result ought to be quite sufficient for our purposes, for if we have no reason to believe that CC is true, then we have no reason to believe that it records a necessary truth, and so no reason to believe that its rejection is incoherent. The only objection to the proposed rejection of CC was that the negation of CC is incoherent; in showing that there is no justification for CC we remove this obstacle, and the invited conclusion - that CC is indeed false - can then be endorsed.

In execution of this strategy, my aim is to identify a necessary condition on any justification of CC. I then show that the indexical argument falsifies this necessary condition, so that given the indexical argument, CC cannot be justified. In short, the indexical argument *removes* all justification for CC, and thus removes all reason to suppose that rejection of CC is incoherent.[2]

Justifying CC

So why do we think that rules are necessary for meaning? (What justification could there be for CC *before* the indexical argument is brought to bear?) Perhaps one reason is that we *feel* like we are following rules when we use language. That is, in describing the world around

---

[2] Some may find the argument here superfluous. Should anyone accept that the indexical argument shows that rule-following is impossible, and that the only apparent way to save meaning is to reject CC, and that this in itself is sufficient to warrant the claim that the rejection of CC is coherent, then I would have no particular objection to raise. However, the onus lies squarely with anyone proposing such an extreme manoeuvre as the elimination of rules from meaning to make the strongest case possible, and here that includes persuading those who think that rejection of CC is incoherent that their view is baseless.

us, we try to follow certain rules, we are rarely in doubt as to how a word should be applied in a new situation, and the activity seems to be as good an instance of rule-following as anything ever does. Yet this phenomenological justification is wholly undermined by the observations made at the end of the previous chapter, that, for the reasons highlighted by Wright, I cannot know that I grasp a rule by introspection, and so am not justified from a first-person point of view in saying that I do grasp a rule. And if I cannot be so justified in saying I *grasp* a rule, I cannot be so justified in saying I *follow* a rule. So as far as one's subjective experience goes, when using language we may feel that a rule is being employed, but that does not indicate that a rule actually is employed, much less that a rule is essential to the activity.[3]

As it turns out, to identify what would justify CC apart from introspection, it is sufficient merely to follow through the consequences of this failure of introspective availability of rules. In particular, if grasp of a rule cannot be known directly by introspection, then even in one's own case, the property (grasping a rule) can only ever be ascribed on the basis of inference. In the case of the third-person, there is nothing unusual in this. To a third party a rule can only ever be ascribed on the basis of their behaviour, in which case the rule is an element within a framework of psychological explanation. With the failure of introspection, though, our knowledge of our own rules must have a similar status: I can have no reason to describe myself as a rule follower other than as part of an explanatory framework.

Of course, in one's own case it is not necessary observe one's own behaviour before describing oneself as a rule-follower, but despite that the distinction between the first- and third-person cases here is negligible. Grasp of a rule is paradigmatically ascribed on the basis that we believe someone is following a rule, that is, that they intend to follow a rule, and that their subsequent actions are informed by that rule. The explanatory power of rules thus arises from their role within belief/desire psychology. To explain the fact that someone continues the series 1002, 1004, 1006..., I need merely cite the fact that she grasps the rule *add 2*, and that she intends to follow it, and that nothing obstructs this aim. In more detail, the explanation for the speaker's utterance '1006' is that she intends to follow the rule add 2, that she *believes* that '1006' accords with the rule in her present context, where the explanation for her belief that '1006' is correct is just that she grasps the rule add 2, and so

---

[3] Wittgenstein suggests that the fact that one's responses come "as a matter of course" is the reason why I myself think that I am following a rule (PI §238). This may be a reason we actually use, but as the argument of the previous chapter shows, it is not a good reason.

knows what the rule requires in any given context. What the rule explains *directly* is the formation of the verdict that '1006' is correct in the present situation. (If she did not grasp the rule, she would not be able to produce the required series of verdicts.) Although a rule is required as part of the explanation of a third party's behaviour, what the rule directly explains is the ability to produce a certain series of verdicts, namely verdicts as to what accords with the rule over varying situations. In both the first- and third-person cases, rules are thus invoked to explain the process of verdict-formation.

A further consequence of the inferential status of rules is that, for it to be at all plausible that rules are constitutive of meaning, then meanings must also be inferential. This follows from the simple observation that a property must have the same epistemic status as its constitutive properties: if F consists in G, and G is inferential, then F must also be inferential. For there to be any hope for a justification for CC, it must be accepted that grasp of meaning fails of introspective availability, and can only be inferred on the basis of explanatory power.

Since both grasp of a rule, and grasp of meaning, are theoretical properties, CC marks a connection between theoretical entities. Consequently, in looking for a justification for CC, we would be as well to ask how constitutive connections between theoretical entities are established in general.

Let us start with an example of an empirical constitutive claim, such as: for something to be an atom, it must contain a nucleus. (The fact that the example is empirical makes it somewhat disanalagous to the justification of the non-empirical CC. The fact that the cases are relevantly different will be accommodated below.) Both the properties *being an atom* and *having a nucleus* are inferential, in that they cannot be established by direct observation. The typical way we should verify such an empirical claim as:

$$\forall x \text{ (x is an atom} \Rightarrow \text{x has a nucleus)}$$

would be to search for a falsifying instance. That is, by testing a wide range of samples, we should see if there is ever something which is an atom without a nucleus.

To simplify matters, suppose that there is a unique test for the existence of an atom, and likewise for the presence of a nucleus. That is, we are warranted in saying that x is an atom if, and only if, we confirm that x has property I_{atom}; and we are warranted in saying that x has

a nucleus if, and only if, we confirm that x has property $I_{nucleus}$.[4] Under these terms, a counterexample to the claim that every atom has a nucleus would arise if we could establish:

$$\exists x\ (I_{atom}x\ \&\ \neg I_{nucleus}x)$$

In this case we should have the right to say that x is an atom, but, having looked at the appropriate evidence, have no grounds to say that it has a nucleus, thus leaving us unjustified in making the constitutive claim that every atom has a nucleus.

Correspondingly, to justify the constitutive connection between atoms and nuclei, we need grounds to dismiss the possibility of any such counterexample. That is, we should have to establish:

$$\forall x\ (I_{atom}x \rightarrow I_{nucleus}x)$$

Unless we have reason to believe that everything which passes the test for being an atom would also pass the test for possession of a nucleus, we would not be justified in making the constitutive connection.

It is not hard to see that this situation should hold quite generally, no matter what theoretical entities are involved. Suppose that confirmation of I_F is necessary and sufficient for the warranted ascription of F, and that confirmation of I_G is necessary and sufficient for the warranted ascription of G. Unless we have reason to believe that everything which is I_F is also I_G, we could not rule out the existence of something which is I_F and not I_G, and so would have no reason to rule out the existence of a counterexample to the relevant constitutive claim. So the general result is that to justify:

$$\forall x\ (Fx \Rightarrow Gx)$$

we must have reason to believe that:

---

[4] In actual fact it may well be that there is no one test for a given property F. (It might be that when x is either G *or* H we can infer than it is F, in which case neither G nor H is *the* property I_F.) For the purposes of the discussion, though, it is acceptable to treat a disjunction of properties as a property, so that I_F is the disjunction of all (strict) properties licensing the inference to the instantiation of F.

$$\exists x\ (I_F x\ \&\ \neg I_G x)$$

is false, for which it is necessary to justify:

$$\forall x\ (I_F x \rightarrow I_G x).$$

To apply this model to rules and meaning, let M be the property of grasping a meaning, and R the property of grasping a rule. The constitutive claim is then:

(CC) $\forall x\ (Mx \Rightarrow Rx)$

And if I_M is the property necessary and sufficient to warrant the ascription of grasp of a meaning, and I_R is the property necessary and sufficient to warrant the ascription of grasp of a rule, then the relevant claim about the relation between evidence for rules and meaning is:

(IC) $\forall x\ (I_M x \rightarrow I_R x)$

As in the general case, unless IC is justified, CC cannot be justified.

So far we have been developing an analogy between the atom/nucleus case, and the rule/meaning case, but they are not wholly analogous, for the former constitutive connection is *a posteriori*, whereas the latter is *a priori*. This difference has two principle effects. In the first place, for CC to be *a priori*, IC must likewise be *a priori*. There is no difficulty with this, though, for if meaning is a type of rule-following, we should expect evidence for meaning to be a type of evidence for rule-following. Mere consideration of various paradigm cases - of when we would ascribe rules/meaning, and the evidence we should accept in each case - ought then be sufficient to establish that IC is true.

The second factor relevant to the difference between the *a priori* and *a posteriori* status of our examples is that in the *a posteriori* case there is a clear justificatory order of priority. In saying that the relevant constitutive claim is empirical, we mean that our judgements about it have to be based on suitable evidence, and in particular on the relation between evidence for the presence of atoms, and evidence for the presence of nuclei. That is, we have to establish that:

$\forall x\ (I_{atom}x \rightarrow I_{nucleus}x)$

before we can establish that:

$\forall x$ (x is an atom $\Rightarrow$ x has a nucleus).

If the constitutive claim under scrutiny is instead *a priori*, there is no obvious reason why this order of priority should not be reversed. That is, suppose that we know that:

(CC) $\forall x\ (Mx \Rightarrow Rx)$.

In virtue of this conditional, anything which counts as evidence for M will *thereby* be evidence for R, so that we would also know:

(IC) $\forall x\ (I_M x \rightarrow I_R x)$.

In this way, the IC is known in virtue of the fact that we know that CC is true. The point is that in saying that the justification of IC is a necessary condition for the justification of CC, we make no comment about the order in which the statements are justified, and it remains open for either to have justificatory priority.

This observation is relevant because the question of whether the rejection of CC is the appropriate response to the indexical argument is at root a question about justificatory priority. To see this, note that the effect of the indexical argument is to show that *nothing* can licence the ascription of a rule. In particular, if grasp of a rule is impossible, then rules have no explanatory value, and so cannot be ascribed on the basis of inference to the best explanation. That is, an immediate result of the indexical argument is that nothing can have the property I_R, or, put more informatively, the result of the indexical argument is that what we took to be evidence of rule-following is not in fact evidence of rule-following. We have to *revise* our opinion as to the identity of property I_R, and given that revision I_R remains uninstantiated.

At this point we are faced with a choice, for there are two opposing directions the argument could take from the above result. One is to take CC as a premise, and from CC establish IC in the way outlined above (i.e. given CC, evidence for meaning will be evidence for rule-

following). Then, since the indexical argument shows that no one in fact instantiates the property I_R, it follows by MTT on:

(IC) $\forall x (I_M x \rightarrow I_R x)$

that no one instantiates property I_M either. Hence we should accept that what we thought was evidence for meaning is not in fact such, thus forcing us to revise our thoughts as to the identity of I_M, leaving it likewise uninstantiated. This results in all meaning ascriptions being unwarranted, and is the route to meaning nihilism.

The alternative direction to take is this. The indexical argument shows that no one instantiates the property I_R, yet gives us no reason to alter our thoughts about who instantiates I_M. In other words, the indexical argument undermines our ability to ascribe rules to people but not meanings. And since people do indeed exhibit evidence of speaking meaningfully (i.e. they instantiate I_M as in so far as we can identify it), but do not instantiate I_R, the *prima facie* conclusion is that IC is false. As established above, the justification of IC is necessary for the justification of CC, in which case it follows that CC is unjustified, and should be rejected. This is the route to rule-elimination.

Put succinctly, the choice here is between using CC to revise our opinion as to the identity of I_M, or to maintain our conception of property I_M, and instead to use this to undermine the justification for IC, and hence to undermine the justification for CC. The question is, which direction is the correct one to take?

It is in answering this question that he order of justificatory priority between CC and IC becomes relevant. For if IC is justified because we know that CC, then in effect we accept that what we think instantiates I_M - and hence what we think property I_M actually is - is determined by CC. And if our identification of I_M *already* rests on CC, we can have no objection if CC is used, in conjunction with the indexical argument, to revise that conception. If, however, the justification of CC rests on IC, then it must also rest on a pre-existent conception of I_M (for we then need to know what I_M is if we are to tell whether IC is true). But if the justification of CC relies on a prior conception of I_M, we cannot use CC to alter that conception.

The dispute is really one between our assertoric practices on the one hand, and our conceptual analyses on the other. If the two are in conflict, which should take precedence? In the abstract, I doubt that there is any principled means of making the decision - whilst we should expect analysis to respect our assertoric practices as far as possible, it can be acceptable for analysis to lead to an alteration of these practices (and presumably it is one of the principle motivations for engaging in analysis that it should do so).

However, in the present case there is no question as to which way to proceed. For the connection between rules and meaning recorded by CC is not something which anyone considers until they come into contact with the rule-following considerations. Although this connection is accepted readily, it is a claim which has to be *appraised*. And indeed, the way we appraise it will be to consider paradigm cases of people who speak meaningfully - that is a situation is which we should not hesitate to say that someone means something - and then to consider whether, in light of this, they are following a rule.

What this means is that it is only against a background of existing assertoric practices that CC can be appraised, and so CC cannot initially inform those practices. We must have some notion of what the property I_M is, and hence who instantiates it, quite independently of our beliefs about the truth or otherwise of CC. It is because in any paradigm case of meaning we should also feel warranted in ascribing a rule that CC is accepted, not the other way around. In other words, whomever we describe as meaning something (i.e. whomever we believe has the property I_M) has to be settled in order to justify CC, and so has to be settled *before* we adopt CC. For this reason CC cannot be used to alter our views as to who instantiates property I_M. Since the indexical argument can only lead us to believe that nothing has I_M (that a meaning ascription is never warranted) *in conjunction* with CC, it follows that the indexical argument cannot be used to undermine the meaning ascriptions that we would otherwise have made.[5]

The key point here is that the justification of CC depends upon our prior assertoric practices with respect to rules. It is only because evidence for meaning coexists with evidence for rules - quite apart from any thoughts as to whether rules are necessary for meaning - that CC

---

[5] It does not follow from the above argument that analysis may never alter our assertoric practices. Rather, the point is that in paradigm cases IC is true, and it is this which makes CC plausible, but this leaves aside the possibility that in marginal cases an analysis may be used to settle difficult categorisations. The point is that the indexical argument has global effect, undermining IC in even paradigm cases, and it is revision of such cases which cannot be supported in the way advocated by the nihilist.

is ever justified. In contrast, since prior to the indexical argument we often have reason to ascribe meanings to people, and since the indexical argument cannot undermine this practice, it follows that we have every reason to believe that people have I_M. That is, people instantiate I_M but not I_R, making conditional IC false. Since the justification of IC was a necessary condition on the justification of CC, it follows that CC is unjustified.

In summary, it is wrong to think that the indexical argument, if unfettered, leads to meaning nihilism. That result only occurs *given* CC, but as we have seen the real consequence of the indexical argument is that CC is unjustified, and so is a candidate for rejection. Rather than trying to save the constitutive relationship by forging some notion of rule-following which is immune to the indexical argument, the solution to our paradox of rule-following is to acknowledge that message that rules-following has nothing to do with meaning.

An Alternative Constitutive Claim

The project so far has been somewhat negative, establishing only that rules are not constitutive of meaning. I now want to put the considerations of the above discussion to a more positive use, by identifying a property to replace rules in an alternative constitutive account of meaning.

The strategy I want to adopt rests on the premise that the only reason we have to reject CC is the indexical argument. That is, I assume that if it were not for the indexical argument, then CC would be fully justified (or at least justifiable). This is a reasonable assumption to make in the present context, for the very value of the rule-following considerations arises from the accepted view that rules are essential for meaning; it is a background assumption of our whole discussion that this claim is more than mere dogma.

On the basis of this assumption, the strategy is to identify what it is that would justify CC were it not for the indexical argument; to thereby formulate a general sufficiency conditions for the justification of a claim like CC; and in this way to find an alternative constitutive claim which satisfies the identified conditions once the indexical argument is brought into consideration.

We need first, then, to identify a sufficiency condition for the justification of CC. To this end, recall that the justification of the conditional:

(IC) $\forall x\,(I_M x \rightarrow I_R x)$.

is a necessary condition for the justification of CC. In fact, IC can readily used to give a sufficient condition as well. Suppose it were indeed the case that I could only ascribe a meaning in cases where I could also ascribe a rule (assuming I have access to the relevant evidence). This relationship is something which ought to be explained, and part of a suitable explanation *may be* that rules are constitutive of meanings. In particular, if meanings give rise, and hence explain the property I_M, and rules give rise to, and hence explain the property I_R, then the fact that a meaning consists (at least in part) in a rule would explain why I_R arises whenever I_M arises. Quite what the best explanation really is would depend very much on what the properties I_M and I_R actually are, but on the assumption that a constitutive relationship is required for the *best* explanation of IC, then on the grounds that we should always adopt the best explanation, we should have a warrant for CC. In sum, a justification of the following two statements:

(IC) $\forall x\,(I_M x \rightarrow I_R x)$

(BE) The best explanation for IC requires that CC be true

is sufficient to justify CC.

Of course the indexical argument entails that IC is not justified, for no one has the property I_R. Yet the above thoughts are in no way specific to rules and meaning, so this framework can be used quite generally. That is, to justify the alternative constitutive claim

(CC°) $\forall x\,(Mx \Rightarrow Fx)$,

namely that meaning consists (in part at least) in F, we should need F to satisfy the following two statements:

(IC°) $\forall x\,(I_M x \rightarrow I_F x)$

(BE°) The best explanation for IC° requires that CC° be true.

In order to identify any such F, the method is to work backwards. To find any such F, we first need to identify some property I_F which satisfies IC°. To find such an I_F, we need to first

identify I_M. And to find I_M we appeal again to the fact that without the indexical argument, the following conditional:

(IC) $\forall x\ (I_M x \rightarrow I_R x)$

is (by assumption) justified.

This fact is useful, for the following reason. Prior to the indexical argument, there is some property I_R which counts as evidence for rules. Given the indexical argument, we find that nothing has the property I_R. But we can still identify that property which *would* count as evidence for rules *were* grasp of a rule possible. Let that property be called PIA-I_R (pre-indexical argument I_R). By definition, anything which was I_R (prior to the indexical argument) is still PIA-I_R (given the indexical argument). The importance of this is that, whereas prior to the indexical argument we would have been justified in stating IC, in the advent of the indexical argument we are still warranted in asserting:

$\forall x\ (I_M x \rightarrow \text{PIA-}I_R x)$.

simply because what was I_R is now PIA-I_R.

It is in this way that what would have counted as evidence for rules can help identify an alternative constitutive property F. For if we can find some inferential property F which satisfies the conditional:

$\forall x\ (\text{PIA-}I_R x \rightarrow I_F x)$.

then given that the conditional:

$\forall x\ (I_M x \rightarrow \text{PIA-}I_R x)$.

is justified (by assumption), then by transitivity we should also be justified in claiming:

$\forall x\ (I_M x \rightarrow I_F x)$.

Once we have identified some such F, it is just a question of seeing whether the alternative constitutive claim is the best explanation for the above conditional. If it is, the alternative constitutive claim CC° would be justified; and if it is not, there may be a better explanation which in turn may justify some other constitutive claim. Clearly the precise result depends very much on the identity of F, which in turn depends on the identity of PIA-I_R, and so a precise characterisation of PIA-I_R - the property which is necessary and sufficient to licence the ascription of a rule, were rule-following possible - is now our immediate concern.

7. Training and Agreement

The initial aim of this chapter is to identify the property PIA-I_R, which (to recall) is defined as satisfying the following:

> Were it not for the indexical argument, confirmation that someone has PIA-I_R would be necessary and sufficient to justify the ascription of a rule to them.

That is, PIA-I_R is what licensed the ascription of a rule before the indexical argument undermined that practice. Once PIA-I_R has been identified, we can begin to formulate an alternative constitutive account of meaning to replace the now defunct one based on rules.

The Explanatory Power of Rules

To identify PIA-I_R, the obvious strategy is to simply ignore (for the moment) the indexical argument, and to ask: what justifies the ascription of a rule? As noted in the previous chapter, a rule can only be ascribed in the context of an explanation for a process of verdict formation. Notably, though, the mere fact that someone produces a series of verdicts under the intention to follow a rule is *not* sufficient grounds on which to infer that a rule is actually being followed. As we have seen (Chapter 5), first-person authority with respect to the grasp of a rule is not to be had, and so the fact that someone thinks they are following a rule is not necessarily an indication that they are indeed following a rule.

As a result, if there is ever a good reason to ascribe a rule to someone (even oneself), it must be in virtue of the verdicts they produce, rather then the fact that they think that they are following a rule. To illustrate the point, note that we should immediately say that someone who produces the series 2, 4, 6,..., is following a rule, even though it is logically possible that this series is either the result of a random selection, or that the agent is 'just disposed' to produce these verdicts.[1] Correspondingly, we should not be so certain that someone who produces the series 2, 6, 8, 17, 56... (which I just made up) is following a rule, even though

---

[1] As was established in Chapter 2, rule-following cannot be reduced to the mere manifestation of a disposition.

we know that any series accords with some rule, so that the production of the latter series above *could* be explained in terms of a rule.

The point is that in deciding whether a given series of verdicts is rule-governed or not, since the subject's own thoughts on the matter indicate nothing of relevance, the only factor we have to base our decision on is the nature of the verdicts actually produced. Given that we are warranted in ascribing rules in some cases but not in others, there must be two different types of series; namely those that we can reasonably expect someone to produce only if they grasp an appropriate rule, and those series we can reasonably expect someone to produce without the aid of a rule. The ability to produce the former type of series is both necessary and sufficient for the ascription of a rule, and would thus constitute the property PIA-I_R.

Clearly what is needed is a criterion which would demarcate the two types of pattern: to distinguish between those series which indicate that a rule is in force, and those which do not.[2] As it transpires, the criterion is this:

> *Rules can be ascribed to those who, as the result of training, attain the ability to produce verdicts which accord with those of their trainer.*

Naturally, this claim needs to be defended. In this respect, given that we are here concerned with training and agreement, and that these two notions have been identified (if contentiously) as essential elements of Wittgenstein's later philosophy of language, it might be thought that Wittgenstein's writings be the obvious place to look for such a defence.[3] However, the multiple layers of exegetical obscurity here already makes the appraisal of the given criterion through Wittgenstein's writings an inefficient strategy. In addition, my own

---

[2] It should be noted that, although rule-following is paradigmatically an intentional activity, the intention to follow a rule is not necessary for the warranted ascription of grasp of a rule. If we can say that someone who intentionally produces a certain type of series must grasp a rule, then likewise someone who unintentionally produces the same series must also be accredited with this property. This is quite unobjectionable, for although someone who (say) adds whilst talking in his sleep may not readily be ascribed with the intention to add - certainly not the conscious intention - we would accept that his behaviour results from his grasp of the addition rule.

[3] Various different community interpretations (i.e. building on some notion of interpersonal agreement) of Wittgenstein's accounts exist (e.g. Kripke 1982, Wright 1980, Malcolm 1989). The rarer claim, that Wittgenstein held that training is essential for rule-following, is made by Fogelin (1987, pp. 175-179), though he goes on to reject it as an untenable thesis. Although many passages in the *Philosophical Investigations* allude to training in connection with rule-following (see for example §§ 197-198), Wittgenstein's most forthright statement about training, given in the *Blue Book* (Wittgenstein 1958, pp. 12-14), actually states that it is a *mistake* to think that the notion of training makes any contribution to the concept of rule-following. (An apparent change in Wittgenstein's opinion in this respect is discussed by Pears 1988, pp. 373-378.)

view is that there is nothing in his writings which could be construed as a convincing argument either for or against the relevant (evidential) claim. Consequently, although we are dealing with markedly Wittgensteinian themes, I shall not attempt to reconstruct any argument of Wittgenstein's as a means of defending the above criterion. The more effective strategy is to construct the argument somewhat afresh.[4]

The aim, then, is to establish that agreement induced by training is both necessary and sufficient to warrant the ascription of a rule. At the outset, it is far from obvious that either the necessary or the sufficiency condition is satisfied. For one thing, although there is little doubt that the subsequent agreement of the pupil with the trainer after a process of training *could* be explained in terms of rules (after all, any series of verdicts could be explained in terms of a rule), it has not yet been established that rule-following is the *best* explanation, that training induced agreement is sufficient to warrant the ascription of a rule on the grounds of inference to the best explanation. In addition, the claim that any kind of agreement - let alone agreement induced by training - is *necessary* to licence the ascription of a rule is strongly counterintuitive. Nevertheless, both conditions do hold, as I shall now show.

Symbols and Practical Success

The best way to establish that training and subsequent agreement are essential for warranted rule-ascriptions is to indicate what is wrong with some other, more immediately plausible, proposals. One idea which has been put forward more than once is that the performance of certain tasks by utilising a series of written marks is symptomatic of a rule-follower. Examples given are: storing, labelling, and relocating things in a storeroom (McGinn 1984, pp. 196-197); following a series of signs as directions allowing one to find one's way (Pears 1988, p. 365 fn 11); and using instructions on a piece of paper to solve a rubics cube

---

[4] Kripke's reading of Wittgenstein gives one explanation for why we should expect no such argument. For if, as Kripke suggests, Wittgenstein gives an irrealist theory of rules, then we should not expect the assertion-conditions for rules to answer to any truth-conditions, and so there is no question of asking what the assertion-conditions ought, rationally, to be. Consequently, all we can do is *describe* those situations which we take to be indicative of rules. In contrast, our aim is to identify the conditions under which such ascriptions *should* be made, to discern not just what people take to be indicative of rule-following, but what phenomenon actually is best explained in terms of rules. The assertion conditions we are after are prescriptive, not descriptive. There is no obvious reason why the descriptive and prescriptive assertion-conditions should coincide, which is why the descriptive project is not our concern.

It might be noted that Kripke's claim that a rule can only be ascribed to someone in the context of a community has met with general incredulity (see for example Blackburn 1984b, McGinn 1984, Boghossian 1989). If Wittgenstein does refer to the community with the aim of describing the conditions under which a rule can be

(Blackburn 1984, p. 297).[5] Certainly such examples have strong intuitive appeal - we do automatically consider such people to be rule-followers - and of course that is precisely the reason such examples are formulated. Quite markedly, these example make no mention of either training or agreement, so that they certainly put some pressure on the criterion recommended above. Yet the issue is not so much whether we would intuitively say that these are examples of rule-following, but to identify those features, common to each example, which warrant the claim that a rule is being followed, and to thereby formulate a criterion of rule-following.

In each of the examples there are two elements at work: (a) the series of actions are goal-directed (for example, several signs are followed in sequence to reach a desired destination); (b) success in the stated aim depends upon the use of a series of written marks. It takes little, though, to see that the inclusion of this second element is wholly disingenuous. The intention is, of course, that the marks be taken as symbols, that they refer to objects in the world, and that this is why they prove to be of (perhaps indispensable) practical benefit. Yet, for a mark to function as a symbol, the agent must grasp a rule which determines what each mark symbolises - a rule correlating an arrow with a particular direction, say. Importantly, the mere fact that the marks are *treated* as symbols - the agent refers to his piece of paper before proceeding in his chosen direction - does not in itself mean that they *are* symbols, only that the agent *thinks* that they are symbols, that he *thinks* that he grasps a rule correlating the marks with objects in the world.

In accepting that the agent uses a set of symbols, we thereby tacitly *assume* that he is a rule-follower. But since we are after conditions which warrant the ascription of a rule in the first place, this is an assumption we ought not make. Rather, if we are to accept that the marks are genuine symbols, then we have to have some reason to say that the agent grasps a rule, a rule in virtue of which the marks refer. And it turns out that any such reason must be given in terms of what the agent does in response to the marks.

---

ascribed, then his answer would appear to be descriptively inadequate. And if the aim is rather to identify the prescriptive project, then clearly the argument (whatever it is) has not been compelling.

[5] Blackburn accredits the example of following instructions to solve a rubics cube to Dummett, but he gives no reference. Each one of the cited examples is given as a counterexample to Kripke's claim (1982, p. 88 ff.) that an individual considered in isolation cannot be deemed a rule-follower, the intention being to give a paradigmatic case of an individual we should say is following a rule.

The point is readily illustrated if we consider someone placing shells on a beach with the aid of a marked piece of paper. The subject uses the paper as he would use instructions; he consults them at each stage before proceeding, but without the paper, he does not attempt any placement of the shells. If the resultant pattern appears to be random, with no organising principle, then we should not be in a position to say that he is a rule-follower: as far as we can tell, this is someone who merely *thinks* that he is following a rule. In this situation we have no reason to suppose that the marks on the paper bear any correlation with anything in the world (to shells, or patterns, or whatever). By way of contrast, if the resultant pattern is the type of pattern which we should expect someone to be able to create only by following a rule, then we should say that a rule is being followed. Since the piece of paper is essential for the agent to be able to carry out the procedure (without it he does not place any shells at all), we have every right to say that the that the marks on the paper do function as symbols, and that the agent is a rule-follower.

As the example illustrates, it is only if we have reason to suppose that the agent is following a rule in his *overall* practice that we have reason to believe that he is following a rule in his use of the marks on the paper. To follow a rule, the agent must correlate his present circumstance with a given action, the action required by the rule. The fact that he can only do this given the piece of paper suggests that the marks function as symbols, that they serve to link input circumstances with output actions in the mind of the agent. The only reason we have to say that they function as symbols is that they facilitate the following of a rule. So we must have some reason to say that the agents in the examples are rule-followers *before* we say that they have a system of symbols. The classification of the marks as a system of symbols rests on the prior application of the desired criterion of rule-following, and so cannot be used in its initial formulation.

As a result, if we have good reason to suppose that our examples are indeed paradigmatic rule-followers, it must be because of the other common element, the fact that the sequence of actions in question is performed in order to meet some overarching goal. To take the example of someone solving the rubics cube, given the large number of permutations involved, we certainly should not expect someone to repeatedly solve the puzzle by chance. The ability to solve it on demand requires an explanation, and the idea that the individual has grasped rules for solving the cube would count as just such an explanation. The suggestion, in general, is that we can call people rule-followers when they achieve an aim which can reasonably be expected to be achieved only by someone who follows a rule, or set of rules.

But, of course, the mere fact that a series of actions is performed in pursuit of an overarching goal does not ensure that a rule is in force. (For example, I might intend to bore someone by reciting a series of random numbers.) What is required, therefore, is a further criterion, to distinguish between those overarching goals which we can reasonably expect only someone following a rule to reach, and those which can be achieved without a rule. Unfortunately, this method does not get us very far, for obviously an overarching aim requires use of a rule only if the series of actions required to achieve the aim itself requires use of an appropriate rule. And that takes us back to where we started, in search of a criterion for a series of actions which only a rule-follower can be expected to perform.

In sum, our 'paradigmatic' rule-followers do not serve to illuminate that which warrants the ascription of a rule in the slightest. In referring to processes which we unhesitatingly accept can only normally be performed by rule-followers - we *know* that only rule-followers follow directions and solve rubics cubes - it may seem that we must be able to ascribe rules in the absence of either agreement or training. But, since it has not been shown why we *ought* to consider these people rule-followers, it has not been shown that our intuitions are right. Such examples do not illuminate the difference between rule-explained and non-rule-explained behaviour, they only appeal to our conviction that there is such a distinction.

Communal Agreement

The above examples were initially designed to show that Kripke is mistaken in thinking that rules cannot be ascribed to individuals in isolation. Given our failure to formulate a criterion of rule-following on the basis of those examples, perhaps Kripke is right after all, and rules can only be ascribed to those whose actions accord with the actions of a community.[6] Indeed, should a group of people behave in a similar manner over a wide range of circumstances, one possible explanation for this would be that they all grasp and follow a common rule.[7] In addition, the existence of such communal agreement would allow us to discount the possibility that each member of the community is simply acting in a random, or in a freely creative way; for the probability of substantial agreement happening by chance is so slight

---

[6] Note that under Kripke's irrealist theory, no justification need be given for why someone can be described as a rule-follower only in the context of a community, whereas in the present context any such communitarian solution must be rationally justified.

[7] By agreement I mean simply that people do the same thing in similar circumstances (calling red things 'red', saying '6' after having said '4'). No element of collusion, or appraisal of others' action is intended.

that we can discount it. So whereas with an individual, any series of acts may be thought to be random, this is not plausible given communal conformity. Since we are trying to find a situation in which the best explanation for a type of behaviour is given in terms of rules, the fact that a certain alternative explanation can be discounted is certainly an advancement.

The trouble with this suggestion is that whereas 'free creation' may be ruled out by communal agreement, there are still many circumstance in which a group of people perform the same actions in correlated circumstances, but in which they are decidedly not following a rule. We all share many dispositions in virtue of a common physical make-up (bones to break when struck, to bleed when cut, and so on). We also share many dispositions due to a similar *psychological* make-up (not to stick one's hand in fire, to eat apples but not stones, etc.). In such cases no rules are required - we behave as we do because of inherent similarities. For example, the fact that we all tend to eat the same types of things does not mean that we have a rule for sorting things into the edible and the inedible; rather, we just accord in our (broadly similar) tastes. Our common beliefs about what is edible do not have to be explained in terms of rules; rather, the disposition to form such beliefs is a basic feature of our 'form of life'. Although this example concerns decisions related closely to physiological matters, there is no *a priori* reason why this should not be the case with any activity - with any belief-formation process - however far removed from basic biological functioning. On any matter on which there is general agreement, a *possible* explanation is that we agree because we have a shared constitution (physical/psychological), and hence have shared dispositions. There is no apparent reason why this type of explanation cannot also be applied when it comes to our common verdicts as to what accords with a rule. If *other* beliefs can be the manifestation of shared dispositions, then why not in this case also?

To be clear: in any situation of agreement, there are two rival explanations on offer. One is that people have a shared disposition to make certain judgements. The other is that people have a shared grasp of the relevant rule. As demonstrated in Chapter 2, grasp of a rule does not consist in a disposition to judge, and so the fact that people have a shared disposition does not entail that a rule is in force. (The difference being that in the latter case the rule provides an explanation of the dispositions under investigation, whereas in the former the dispositions are taken as a basic fact.) If we are prepared to give an explanation in terms of dispositions in the case of an individual, then there is no objection to the thought that we share these dispositions; after all, we share the same physiology, psychology, and social upbringing.

Admittedly, there are situations in which an explanation of communal agreement in terms of basic common dispositions (rather than rules) would put severe strain on our intuitions. The matter is particularly acute when we consider examples of actions performed under the common intention to follow a rule, and hence where all the participants believe they are following a rule. Should a group of people spontaneously intend to follow a rule, and happen to broadly agree in their verdicts - suppose that there is a community of natural-born mathematicians, for example - it is difficult to resist the claim that such people do have an inherent grasp of the relevant rules.

However, the mere fact that the members of the community all think that they are following a rule is not in itself a reason to suppose that a rule is indeed being followed. For, since an individual may think he is following a rule when in fact he is not, it is likewise possible that every member of the community is similarly afflicted. In support of this claim, it might be noted that whilst in some situations the existence of non-collusive communal agreement makes error less likely (the thought being that many pairs of eyes, say, are often better than one), this is not a useful observation in present circumstances. If someone falsely believes that they are following a rule, their mistake will not be the result of perceptual error. Rather, any error is likely to be either systematic (in that we have false view of what counts as evidence of rule-following), or blameless (in that what is good evidence for rule-following has arisen, making our judgement fully justified, when, as a contingent fact, a rule is not being followed). Under either scenario, the presence of communal agreement offers no advantage: there is no reason why the false view as to what counts as evidence for rules should not be widespread. And if evidence for rule-following happens to arise in the absence of a rule, we are all likely to be taken in. So it would be a mistake to regard the fact that everyone in a community believes that a rule is in force as a better indication that one is actually in force than it is in the case of an individual.

The result is that when faced with natural-born adders, as with the example of those using written signs, the overriding intuition is that rule-following must be involved, that the situation cannot reflect the mere possession of a disposition. On investigation, though, we find no rational foundation for this preference. But again, rather than merely recording our differing sentiments, we need to know why one explanation should be preferred in some cases and not in others. And the fact that communal agreement does not automatically

commend an explanation in terms of rules means that there must be other factors at work which we have not yet identified.

Training

We are now in a position to see why training is an essential for the warranted ascription of a rule. The most significant new element that training brings to the picture is mention of the origin of the putative rule-follower's behaviour. Notably, a (successful) training process results in the trainee acting differently from the way he would have acted otherwise. Prior to induction in the rule the trainee cannot accord with the trainer in new cases; after the training he can. This opens up a new explanatory need, for whereas the existence of dispositions may not always stand in need of explanation (a person's dispositions being a basic fact about their constitution), a *change* in someone's dispositions ought to be accounted for. In particular, we should want to know how a process of training produces agreement when before there was none.

To bring the matter into focus, consider a typical situation in which H teaches Johnny how to count. In her role as trainer, H will herself follow the rule a number of times by way of demonstration, then get Johnny to try to continue the pattern, and either praise or correct the responses he gives. At first Johnny's judgements may be tentative, but with his increasing level of success he gains more and more confidence, until at some stage (assuming the training is successful) Johnny will be able to reproduce H's answers quite readily, without a moment's hesitation. We should then say that Johnny has learnt the rule. Before the training, Johnny could not have produced the same answers that H produces, but after the training he can. It is this change to Johnny that we seek to explain, bearing in mind that it is quite typical, that a similar process of training can be expected to produce a similar change in any normal person.

It is clear that an explanation in terms of rules is perfectly adequate to account for this change. On the basis of the training, Johnny grasps the same rule as H, and in virtue of their respective grasp of the same rule, each knows (*ceteris paribus*) what the rule requires. The only obvious deficiency with this type of explanation arises from the need to account for how a finite process of training succeeds in transmitting the intended rule. Training involves the demonstration of only a limited number of applications of a rule, and as is well established, any finite sequence is consistent with an infinite number of rules. In purely probabilistic terms, then, there is no chance that the trainee will arrive at the intended rule,

the one actually being followed, given the sheer number of rules consistent with the evidence.

To solve this puzzle, we should first note that the trainee does not recognise that there is an infinite number of rules each of which is consistent with the trainer's behaviour. In fact, at the outset there can be *no* rule that Johnny grasps which is consistent with H's behaviour. This is because, were Johnny to grasp some rule consistent with H's behaviour, he would see her behaviour as a manifestation of that rule, and not the inculcation of a new one. It is not then a wholly rational process which leads Johnny to give the answers he does - he does not infer that the trainee intends one rule rather than another, for he lacks the conceptual resources to make that particular inference. And if the responses Johnny gives are not the result of a rational hypothesis about which rule is being taught, anyone advocating an explanation of Johnny's behaviour in terms of rules has little option but to accept that training *causes* grasp of a rule. It is then quite plausible that similar training processes produce the same result in similar subjects. Given that the trainer and trainee are in many ways alike, it is no accident that the same type of process which instilled a specific rule in H will instil the very same rule in Johnny.

As might be expected, in accounting for the whole process of training and subsequent agreement, the explanation which refers to rules is not the only option. An alternative proposal is that the training causes Johnny to be (merely) *disposed* to give answers which happen to be the same as his teacher's. Such an explanation does not refer to rules, for the disposition to judge does not constitute grasp of a rule. Yet this type of explanation is also adequate in accounting for subsequent agreement between trainee and trainer. On the assumption that the various training processes undertaken by different individuals bear some similarities to each other, and given the similar physical/psychological properties of their trainees, we can expect the same training process to yield similar dispositions. On the assumption that the training received by the trainer is similar to the training she gives to others, we could reasonably expect her pupils to acquire dispositions which accord with her own.

Again, the intuitive view is certainly that here an explanation in terms of rules is preferable to the given alternative in terms of dispositions. Yet the purely dispositional explanation ought not be discounted out of hand, for causal explanations of a change to one's psychological dispositions - including the disposition to judge - are quite acceptable. For

example, a 'truth serum' may make people more disposed to tell the truth; alcohol makes people disposed to misjudge the width of their cars; and hypnotists can make people believe anything at all. Given that causal explanations for dispositions to judge are applicable in other contexts, it is necessary to provide a reason why such explanations are *not* to be applied to situations of training induced agreement as to the requirements of a rule. If a rule explanation really is superior to a dispositional explanation in such cases, there must be some basis to this differentiation.

To see why the rule-explanation is the best explanation of the training process, it is necessary to identify what is to be explained more precisely. In the initial stages, Johnny is presented with a series of examples of how the rule is applied, and is then invited to say how the rule should be applied next. The most important point to note about this situation is that at each stage Johnny forms his opinion as to how H's rule should be next applied *on the basis of H's behaviour*. That is, Johnny judges that φ comes next on the basis of what he has been shown so far - on the basis of H's demonstrations, and the appraisals H has made of Johnny's previous attempts at following the rule. So when the training has been deemed a success, Johnny not only forms verdicts he would not previously have made (for example that φ comes next), he does so *for reasons* which he would not previously have accepted. Explicitly, he now takes H's behaviour to indicate that φ comes next, when previously he would not have. What is to be explained, therefore, is not only why Johnny forms new verdicts, but why he comes to adopt what he does as his reasons for those verdicts.

At the outset, we are still faced with two competing explanations, namely:

Rule explanation: The training causes Johnny to grasp a rule, and the rule explains why he accepts such-and-such as a reason to believe that φ is correct.

Disposition explanation: The training causes Johnny to (be disposed to) accept such-and-such as a reason to believe that φ is correct.

Both are perfectly adequate explanations for the change which Johnny undergoes during training. Yet, as we shall see, only the rule-explanation accommodates the rationality of the subject, and this is enough to make it the better explanation.

Reasons and Causes

The superiority of the rule-explanation is established by comparing three types of explanation for belief. Suppose we have a situation in which:

(a) Mary believes that p

(b) Mary believes that p for reason x.

One way to explain Mary's belief would be simply to cite reason x: Mary believes that p because x. This would be a *reason explanation*.

Suppose, though, that Mary only believes that p because she received a sharp blow to the head. However, Mary does not recall being hit on the head, and when asked why she believes that p, she still cites x as her reason. Since she would not have held x to be a reason to believe that p prior to the blow (for she was aware of x, but yet did not believe that p), it seems that the blow *caused* Mary to adopt x as a reason to believe that p. The situation is captured thus:

(c) Mary believes that p

(d) The reason for Mary's belief that p is x.

(e) The cause of Mary's acceptance of x as a reason to believe that p is y.

In this case, the explanation of Mary's belief is a *caused-reason* explanation. The belief in question is still held for a reason, but now the cause of the adoption of that reason as a reason is a relevant factor in the explanation.

It is the distinction between a caused-reason explanation and a reason explanation which is of interest, but it is worth introducing a further distinction, concerning a special type of caused-reason explanation. Turning to another example, suppose that a doctor gives Mary ecstasy for a quite legitimate medical reason. On taking the drug Mary becomes paranoid, and suffers delusions of persecution, making her disposed to misinterpret the actions of others as being acts of hostility directed against her. As a result, Mary takes the fact that the doctor administered the drug to her to be evidence that he (the doctor) intended her harm. In this situation the doctor's administration of the drug is *both* the (albeit irrational) reason for Mary's belief that the doctor is against her, *and* the cause of her acceptance of it as a reason. This situation is captured as follows:

(f) Mary believes that p

(g) The reason for Mary's belief that p is x.

(h) The cause of Mary's acceptance of x as a reason to believe that p is *also* x.

In this case, where the same state of affairs is both the reason for a given belief *and* the cause of the subject's acceptance of it as a reason, the appropriate type of explanation for Mary's belief may be termed a *self-caused reason* explanation.[8]

This latter type of explanation is pertinent to our discussion, for one of the accounts of training induced agreement we are assessing - namely the dispositional account - is in fact a self-caused reason explanation. As already noted, this type of explanation accepts:

(i) Johnny believes that φ comes next

(j) The reason for Johnny's belief that φ comes next is H's behaviour

(k) The cause of Johnny's acceptance that H's behaviour is a reason to believe that φ comes next is H's behaviour.

This fits the schema for a self-caused reason explanation.

The fact that the dispositional account is specifically a self-caused reason explanation is not overly significant. What does matter is that the dispositional account, *qua* self-caused reason explanation, is a *type* of caused-reason explanation; and that there is a clear criterion for when a reason explanation is preferable to a caused-reason explanation.

To uncover this criterion, first note that a caused-reason explanation (indeed, a self-caused reason explanation) could, in principle, be offered for *any* belief formed for what are ostensibly rational reasons. For example, suppose that Jim says to me "Your house is on fire". As a result, I come to believe that my house is on fire, and the reason for my belief is what Jim said. An orthodox explanation - a straightforward reason explanation - for my

---

[8] In the discussion above I treat a reason for belief as a worldly state of affairs, whereas more usually a reason for a belief is considered to be some other belief. This fact can readily be accommodated within the account: a self-caused reason would then involve a worldly state of affairs X which causes me to take *my belief* that X to be a reason for some further belief that Y. The central point used in the argument - that such cases involve an alteration of one's rational processes - is not affected if we ignore this distinction, but doing so makes the presentation easier.

belief would advert to various facts about my overall view of the word, as well as to my inferential practices. In particular, we should note that I believe that Jim speaks English, that he is a reliable witness, and that as a result of these beliefs I conclude that my house actually is on fire. An alternative explanation would be: Jim telling me that my house is on fire *caused* me to accept his telling me that my house is on fire as a reason to believe that my house is on fire. This explanation fits the model of a self-caused reason: the same situation is both my reason for belief, and the cause of my adoption of it as a reason.

In most situations we should prefer the reason explanation to the self-caused reason explanation, even though each is explanatorily adequate. The reason for this emerges when we examine clear-cut cases of caused-reason explanations. Such examples are readily found: for example the situation in which Mary is hit on the head, but does not remember it; or when she is injected with ecstasy, and becomes paranoid, but without her being aware that a drug has been administered. In these case, the cause cannot be the reason for belief, simply because the subject is not aware of the occurrence of the relevant event.

What is notable about all examples of clear-cut caused reasons is that there is no expectation that the ensuing beliefs should meet accepted standards of rationality. Thinking of situations in which someone receives a blow to the head, is administered psychoactive drugs, or is a subject of hypnoses, it is possible that any belief, no matter how bizarre, be the outcome of such procedures. Of course, it might be that someone is hypnotised, and so caused to believe something perfectly rational which they would not otherwise have believed. But any such accordance with rational standards would be entirely accidental, for it is just as likely that these events produce wholly irrational beliefs. Indeed, it is only when the subject's otherwise rational belief-formation process is *disrupted* that we can say for certain that the reason is adopted as a result of the relevant event, rather than as the result of their existing belief-formation processes.

In contrast, it is when exercising reason that one sets one's expected standard of rationality, and so when exercising one's rational faculties, we have every right to expect that rational standards be met. And it is in precisely this situation, in describing the use of one's rational faculties, that a reason explanation is appropriate.

The general result is that if one's belief-formation processes do accord with broadly rational standards, and we have no reason to believe that this is merely accidental, then a causal-

reason explanation is to be rejected in favour of a reason explanation.[9] Applying this to our trainee, it follows that if in responding to training Johnny meets certain rational standards, and this is not accidental, then an explanation for his change in behaviour in terms of (self-) caused reasons is inadequate. All that would then be needed to establish the superiority of the rule explanation would be to show that it, in contrast to the dispositional explanation, can account for the rationality of Johnny's responses.

The Rational Trainee

There are, then, two questions before us. One: when responding to training, does Johnny act rationally? Two: if so, can a rule explanation of the change which Johnny undergoes during training account for this rationality? Let us look at these questions in order.

In asking whether a trainee acts rationally, what we specifically want to know is what happens when Johnny, having been exposed to a limited fragment of a rule (perhaps accompanied by some initial corrections of early mistakes and praise for a few correct attempts at extrapolation), now decides that φ comes next. Our question refers to the pedigree of this belief: is it the product of a rational belief-formation process?

To raise some initial doubts that a trainee does act rationally, it is certain that Johnny may respond to training without any of his actions being *fully* justified. For one thing, to respond to training Johnny must believe that H is following some rule which she is trying to teach him, otherwise he would not attempt to continue the rule in new cases, and not attempt to 'latch on' to the rule. Yet Johnny need initially have no concrete evidence which would indicate that H is a rule-follower. As noted above, the fact that someone performs a series of actions, even under the belief that they themselves are following a rule, is not in itself a good reason to suppose that they are indeed following a rule. Yet this is precisely the situation that Johnny may find himself in - he is faced with H who manifests a certain range of behaviour, and on this basis alone Johnny must come to believe that she is following a rule. It may well be that in order for the actions of H to make any sense to Johnny, he first *hypothesises* that H is following a rule, which is to say that he adopts an unwarranted assumption.

---

[9] Saying that someone satisfies certain standards of rationality is not to credit them with grasp of a rule for rational belief formation (a situation which would make the search for a criterion of rule-following circular); to merely accord with a rule does not mean that a rule is grasped, or followed, or in any way 'in force'.

A second reason why Johnny's actions may not be fully justified is that any series of verdicts accords with some rule. But if any verdict which Johnny gives is consistent with the preceding finite behaviour of H (under the hypothesis that H is following a rule), then the verdict he chooses to give is as well supported as any other, which is to say that his choice is strictly unjustified.

Neither of these observations precludes the possibility of Johnny acting rationally, for we are not interested in absolute standards of rationality. Rather, we want to know whether his belief is rational *given* his existing background beliefs. In particular, no one would respond to training unless they accepted that a rule was being demonstrated to them. So, in accepting Johnny as a trainee, we require that he already believes that H is trying to teach him a rule. This belief is an initial condition, and requires no justification.

There is another background belief which anyone learning a rule must also utilise, namely the belief that a finite fragment of a series can determine a rule to within uniqueness. In light of our investigations we know this to be false, but it is generally accepted amongst the population at large. Indeed, this assumption underlies one paradigmatic test of rationality, the ubiquitous IQ test question: "What is the next number in the series...". This type of question relies on the fact that people do accept that there is one unique answer, that the finite fragment does admit of only one expansion, and that we have been given sufficient information to identify it.[10]

With these background beliefs in place, the process of training proceeds as follows. To begin with, H demonstrates a fragment of the rule, after which Johnny is invited to continue it in a new case. At this stage Johnny may have no firm belief as to what comes next - his answer may be tentative, his aim being to discover whether his answer is correct. Should his answer be endorsed, he will try to continue to apply the rule in further new cases; upon correction of a wrong answer, he will try to take this new information into consideration, and try something else. If the training is eventually successful, Johnny will meet with repeated success, and will gain confidence in his answers. He then changes from testing tentative answers with his trainer to given assured responses.

---

[10] Once the Wittgensteinian point has been made we can still infer what the intended rule is because we know what tacit principles people adopt when they formulate such questions. The key point, though, is that in everyday cases we do not consider that there are possible rules which we fail to take into consideration.

What is characteristic of this process is that it is very much like one of hypothesis testing. The trainee acts *as if* he hypothesises that the trainer is following some specific rule: continued agreement acts as confirmation, thus adding to the trainee's confidence that he has got it right; dissonance serves to falsify the hypothesis, in which event the trainee rejects it, and so starts to test some other hypothesis in its place.

Given the background beliefs that the trainer is following a rule, and that the identity of this rule can be determined from a finite number of applications, a genuine process of hypothesis testing following this model would be an entirely rational process. In analysing the process of training, though, we ought not take this talk of conjecture and refutation too literally. In actually learning a rule, we do not engage in a genuine process of hypothesis testing, for as noted a trainee does not yet grasp the rules necessary to make an explicit hypothesis as to the rule being demonstrated.

Nevertheless, to meet standards of rationality, all that is needed is that people act 'as if' engaged in a process of hypothesis testing. In particular, in training it is essential that we respond to praise and correction respectively in the appropriate manner. No matter what verdicts it leads me to suggest next, I must take note of this ongoing appraisal, change what I was doing if it is not working, and build on answers which have met with approval. It is this process - the retention of approved responses, rejection of censured responses, and consideration of both types of fresh information when giving further responses - which makes the whole process emphatically rational. On this basis we can discount the self-caused reason explanation of the training process, for that explanation does not accommodate non-accidental rationality.

The Rule Explanation

It remains to be shown that the rule-explanation of training is consistent with the rationality of the trainee. This does not mean that the acquisition of a rule must be *purely* the result of ratiocination, only that whatever processes are involved prove to be consistent with the application of rational belief formation processes. As mentioned above, in explaining Johnny's behaviour in terms of a rule it is accepted that the training causes him to grasp the rule, so there is still an element of causation within the account. The important point is that although there is a causal element within the account, this feature does not *cause* Johnny to accept the behaviour of his teacher as reason for his belief. Rather, by utilising his pre-existing inferential practices, the caused change *enables* Johnny to see his teacher's

behaviour as a reason for his belief that φ comes next. It is this causal change which allows Johnny to make an inference (i.e. that φ comes next) that he would not otherwise have made.

Again, the situation can be likened to hypothesis testing. To give an analogous situation, suppose that the apple falling on Newton's head caused him to acquire the concept of a force, and so to be able to formulate and entertain the theory of universal gravitation. If so, he could subsequently reason with his newly acquired concept, is in the position to test a new hypothesis, and so might be able to reach fully warranted conclusions which he could not have rationally reached before. Importantly, the scope of his rationality would thus be expanded, but the nature of his rational belief-formation processes would remain intact. The conclusion that the motion the planets is governed by gravity could then be justified quite rationally. In contrast, if the apple falling on Newton's head caused him to believe something he did not believe before (but for which he did have the necessary concepts), or to accept something as a reason for a belief which he did not previously accept, then Newton would not then be employing his rational belief-formation processes, and the whole procedure is not rational.

Likewise with a rule-follower: in coming to grasp a new rule, the type of behaviour Johnny 'recognises' as being rule-governed is altered. That is, someone who grasps the rule add 2 will recognise the series 2, 4, 6... as a fragment of the series generated by adding 2, and so unhesitatingly forward 8 as the next element of the series. His rational processes do not change, but he can now 'see' various options for continuing the series in ways not previously available to him, and as a result he can reach a new rationalistic conclusion as to what comes next. Although there is a causal element in the explanation for the change to the inference he makes - he is caused to grasp the rule - the explanation is not *wholly* causal, and does not conflict with our overall view of him as a rational agent.

In summary, it is the following features which make training induced agreement necessary and sufficient for the warranted ascription of a rule. First, what is to be explained is the origin of a certain mode of behaviour on the part of the trainee. Second, the change in behaviour is a change which is characterised by an on-going agreement with the responses of the trainer. Third, training occurs when the trainee puts himself under the authority of the trainee, so that his aim is to state how the rule continues based on the evidence of the trainer's prior behaviour. The trainee's judgements as to what comes next are thus always made for a reason. These features of training leave open two possible explanations: one

directly causal (the trainee is caused to reason differently), one indirectly causal (the trainee is caused to grasp a rule, and this explains why he reasons differently). Fourth, in responding to training the subject must reject/endorse answers as indicated by the trainer and use the information when forming new verdicts. In other words, the trainee must be a rational agent. Fifth, to explain a process which satisfies accepted standards of rationality, a reason explanation is superior to a caused-reason explanation. And finally sixth, the explanation in terms of grasp of a rule is a type of reason explanation. It is only when the appropriate change occurs within a framework of rationality that the rule explanation is markedly better than the alternative, and since the rule explanation is wholly adequate, the situation is necessary and sufficient to warrant explanation in terms of rules. The ability to agree as the result of training is, therefore, the required property PIA-I_R.[11]

Meaning as Communal Use

Our reason for identifying PIA-I_R was to motivate an alternative basis for meaning. Given some property F such that:

(IC°) $\forall x\ (I_M x \rightarrow I_F x)$

(BE°) The best explanation for IC° requires that F be constitutive of meaning,

then we should have established:

(CC°) $\forall x\ (Mx \Rightarrow Fx)$,

i.e. that F is constitutive of meaning. In pursuit of some such property, we noted in the previous chapter that should some property F satisfy:

(*) $\forall x\ (\text{PIA-}I_R x \rightarrow I_F x)$

---

[11] Earlier it was noted that in some situations we should intuitively describe people as rule-followers even though they have not undergone any training. (The example considered was of people who are born mathematicians.) The basis for our inclination to describe such people as rule followers could be given thus: our born mathematician behave in a similar manner to ourselves (adding, subtracting, and so forth), and given that we explain our own behaviour in terms of rules, we should apply the same explanation to our born mathematicians on inductive grounds. Once, however, it is recognised that training is instrumental in the justification of the initial ascription of a rule, the inductive basis for this inference breaks down, making the intuitive description unwarranted.

(i.e. what would be evidence for rules is also evidence for F), then F will automatically satisfy condition IC°. In this way, *using* PIA-I_R to identify properties which satisfy IC°, we may come up with a range of properties, each of which is at least a candidate for a constituent of meaning. It would then merely be a matter of seeing which, if any, of these candidates also satisfies BE°. If one does, then our task is complete.

So, which properties satisfy (*)? Given that PIA-I_R is the ability to agree with one's trainer having undergone training, any such F would be ascribable to anyone who manifests that ability. One way we could ascribe a property in this situation would be if F *explained* that ability (F thus being ascribed on the basis of inference to the best explanation). It is by this route that the constitutive relation between rules and meaning would have been established, for a rule is, as we have seen, just the type of thing which could explain training induced agreement (that is, until the indexical argument shows that this type of explanation is unavailable). Nothing else comes to mind which could take on this explanatory mantle - what could explain why we all think a given action is correct if not a rule? - which leaves the ability a basic property, not to be explained in terms of anything else.

As a result, there is no option but to accept that F can be ascribed to anyone who has the ability to agree with others as the result of training because the property F *is* the property of possessing this ability.[12] The explanation for the fact that that meanings can only be ascribed to those who manifest training induced agreement is that meaning consists in this ability. In as much as we thought that the truth-rule for a word were necessary for its meaning, it turns out that we are as justified in thinking that the ability to use it in accordance with our teachers - to agree in our verdicts as to what is correct - is constitutive of meaning.[13] We do

---

[12] In talking of an ability to agree with the community, the reference to an ability needs careful handling. We might analyse abilities in terms of dispositions - so that the ability to ride a bike consists in the disposition to do so given both the intention to ride a bike, and the appropriate means. In the present case, though, the ability to agree does not mean that we would agree given the intention to agree: the aim is not to agree, but to act correctly, or to speak the truth. In this case, the intention is askew from that of a normal ability, but the description of it as an ability still appears apt.

[13] The theory may be considered an explicit answer to the worry expressed by McDowell, who, when criticising Wright's (1980) communitarian account of rule-following, says:

> The problem for Wright is to distinguish the position that he attributes to Wittgenstein from one according to which the possibility of going out of step with our fellows gives us the *illusion* of being subject to norms, and consequently the *illusion* of entertaining and expressing meanings. (McDowell 1984, p. 336)

The point made here is precisely that the 'illusion of norms' created by collective on-going agreement, far from being an illusion of content, is all that content actually requires.

not have to follow rules in order to speak meaningfully. Rather, we merely have to *appear* to follow rules.

Communal Use and the Indexical Argument

To complete the argument for this, what we may call the training-induced communal use theory of meaning (or *communal use theory* for short) a final step is needed.[14] We have already noted that once rules are removed from consideration, the indexical argument poses no threat to meaning; and on this basis it is *automatic* that the communal use theory does not itself succumb to the difficulties raised by the indexical argument. Nevertheless, since the principle recommendation for the communal use thesis is that its competitor (rule-based) theory is precluded by the indexical argument, to make the communal use theory fully secure it will help to show precisely *how* this immunity to the indexical attack is achieved.

The question underlying the indexical argument is, of course, this: what determines whether I mean the same now as I did previously? Under the communal use theory, the answer is unsurprisingly that one means the same just in case in using a word, one engages in the same communal practice as before. This answer, though, is not wholly satisfactory, for it immediately raises the further question: what determines that we are now engaged in the same practice as before?

Here the answer comes in two stages. The first point is that under the communal use theory, nothing as yet determines how a community ought to behave if it is to continue a practice in the future. Rather, the use of a word will be fixed by the community at the time. In that case it looks as though any present communal behaviour could be a continuation of any previous practice, in which case we would have the indexical problem all over again. But we have so far omitted the vital point that the practice in question must be the result of training, which for each individual is a determinate historical process. In that case, the community continues the same practice on condition that the behaviour of each member arises from the same historical training process. So indeed, although in principle any practice could be the continuation of some earlier one, in actual fact it will only be such a continuation if both emanate from the same source. It is because of the role training plays in the communal use

---

[14] The reference to the community here merely signals the role of interpersonal agreement between trainee and trainer within the account.

thesis that a communal process, unlike a rule, has determinate trans-contextual identity conditions.[15] The communal use theory does indeed neutralise the indexical argument.[16]

---

[15] To be clear, it is no part of the present thesis that communal verdicts determine the extensions of our words, and so this is assuredly not a version of communal rule-following.

[16] It is worth confirming that the elimination of rules avoids the objection made against Kripke's 'sceptical' solution. In Chapter 4 it was argued that meaning irrealism is incoherent on the following basis: meaning irrealism is an essentially explanatory theory (prescriptive explanation); irrealism of any kind strips its subject matter of its (prescriptive) explanatory power; that meaning irrealism is a self-applicable thesis; and that as a result meaning irrealism strips itself of its required explanatory power. Although the use theory here defended is also self-applicable (as any theory of meaning must be), it is distinctive in that it explicitly forgoes the role of prescriptive explanation. Without norms, there can be no account of what people ought to say because there is nothing that people ought to say. In particular, whereas the assertion-conditional theory is designed to explain how attributions of rules are still prescribed even in the face of the 'sceptical' argument, the communal use theory denies meaning ascriptions (like any other ascription) are normatively required. For this reason the same incoherence does not arise.

8. Meaning, Use and Truth

The notion of meaning is so intimately connected with that of truth that we cannot be said to have a measure of the communal use thesis unless we have some idea of its attendant theory of truth. Indeed, a theory of language will likely only have philosophical ramifications in so far as it tells us about the language-world relationship, and we can hardly say anything informative here without mentioning truth. So, under the communal use theory, what does happen to truth?

The communal use theory has been established on the basis that rule-following is impossible, which is to say that all notion of correctness has been relinquished. The *prima facie* conclusion to draw is that since truth is a type of correctness, meaning does not determine truth at all; that no linguistic entity has a truth-condition, and no statement or utterance ever has the property of being true, or of expressing a truth. This 'no truth' theory of truth is admittedly an extreme position, carrying with it little inherent plausibility, and indeed it is a position of dubious coherence. (If there is no truth, the 'no truth' thesis cannot itself be true.)

There is, though, an altogether more sophisticated approach to the connection between meaning and use which ought to be considered in the present context. The thesis, proposed by Horwich (1995), is based on a version of deflationary truth, Horwich's version being that the concept of truth is given by the disquotational schema:

(DS) 'p' is true $\Leftrightarrow$ p

In saying the DS captures the notion of truth, Horwich means that all the properties of truth, the explanatory power of truth, and grasp of the concept of truth, can all be accounted for in terms of DS.[1]

---

[1] Various characterisations of deflationism are available. I focus on Horwich's account initially, but consider how different formulations stand with respect to the argument below. Note that Horwich uses DS to characterise truth in the paper I discuss here (Horwich 1995), but that in his (1990) he gives a different version (namely that truth is captured by schema ES given below). As discussed in the main text below, DS is certainly inadequate as the basis for a theory of truth, but I defer the discussion of its flaws to allow the principles at work to be exposed quickly.

Horwich's central contention is that apparently non-normative use is in fact normative, and can determine truth-conditions, on condition that truth is deflationary. Indeed, in this way Horwich re-invigorates rule-following, for use is claimed to determine a correctness condition (under a correspondingly deflationary notion of correctness) after all. Nevertheless, the thesis in no way conflicts with the foregoing argument that rule-following is impossible, for (as shall become clear) that argument was directed against a robust (i.e. non-deflationary) theory of correctness. If Horwich's thesis turns out to be right, then we need not reject the earlier argument, but merely recognise that it establishes the impossibility of robust rule-following, a result which motivates deflationary rule-following.

Before assessing the compatibility of the communal use theory and deflationary truth, it is important to note that Horwich's concerns about the relationship between deflationary truth and use are slightly different from our own in three respects. First, Horwich accepts that truth-determination is a criterion on meaning; that a use theory cannot be endorsed unless it allows for truth-determination. In the present context our project is somewhat more investigative: having established the communal use theory, our task is to identify the consequences for truth, and to ensure that the result is coherent.

The second minor difference in Horwich's position is that in talking of the use theory, he means simply that meaning is dispositional, without any mention of either training or the community. As we shall see towards the end of this chapter, this difference does have some bearing on the cogency of Horwich's argument. However, the initial prospect is that under deflationism, *any* meaning-determining property (and so certainly any version of a 'use' theory) can determine truth. To avoid prejudice against Horwich, I shall initially consider the argument as it is presented, and only then look at how the different ways 'use' is employed have any bearing.[2]

The third difference is that Horwich accepts, as a premise, that the deflationary theory is correct. I shall not here consider independent arguments either for or against the deflationary

---

[2] Field (1994a, 1994b) defends a similar position to Horwich, claiming that deflationary truth is compatible with a 'use' theory of meaning (although Field's definition of deflationism is different from Horwich's). It is notable that Field takes the 'use' theory to be inherently normative - 'use' involves norms of warranted assertion - whereas Horwich appeals to deflationism to *provide* use with its normativity, so that use is normative in a deflationary sense. Clearly any theory which presupposes norms of warranted assertablity cannot be endorsed in the present context without begging the question.

theory,[3] but merely ask whether it is *consistent* with a use theory of meaning. This is because *by default* any remotely plausible notion of truth will be better than no truth at all (and deflationism is at least initially credible), so that mere consistency with the use theory would in itself be a sufficient recommendation for deflationism.

Horwich's Deflationary Theory

The key feature of deflationism which Horwich's exploits is its claim that "truth is not susceptible to conceptual analysis, and has no underlying nature" (Horwich 1995, p. 358). This is a direct consequence of the given characterisation of truth. For comparison, a more usual (non-deflationary) theory of truth would be of the form:

'p' is true $\Leftrightarrow$ 'p' is F

Such a theory presents an analysis of the concept *true*. As Horwich notes (1995, p. 360), many such theories have been offered, but so far without widespread acceptance. The deflationary theory is signally not of the above form, and entails no statement of this form, which is why Horwich states that truth, under deflationism, is not susceptible to conceptual analysis.

The relevance of this distinctive feature of deflationary truth to the use theory of meaning is demonstrated in Horwich's diagnosis of Kripke's rejection of the dispositional theory of meaning. To recall the discussion of Chapter 1, Kripke objects that dispositions cannot determine truth on the basis that there is no set of ideal conditions O such that:

S is true $\Leftrightarrow$ S would be uttered in conditions O

This has the form of a traditional analysis of truth, which means that in adopting the satisfaction of this conditional as a criterion on a dispositional theory of meaning, Kripke thereby assumes that truth must be susceptible to analysis. In doing so he assumes that truth is robust. So, Horwich contends, rather than showing that dispositions are not truth-determining, all Kripke actually proves is that dispositions cannot determine *robust* truth. The possibility remains that dispositions can determine truth, if truth is deflationary.

---

[3] Horwich's theory is developed and defended comprehensively in his (1990).

Kripke's assumption (that truth must be analysed in terms of dispositions for the dispositional theory to succeed) is quite understandable given the way reductions normally work. For example, to say that water reduces to H_2O, we should first have to show that all the properties of water - its transparency, fluidity, boiling point, and so on - are properties of H_2O. Similarly, before reducing meaning to dispositions, we should show that the properties of meaning - including truth-determination - are properties of a dispositional state. Since truth-determination is not an empirically verifiable property, the only evident means of *showing* that dispositions determine truth is to give an *analysis* of truth in terms of dispositions.

In setting himself in opposition to this traditional approach, Horwich must show how use may determine truth without the benefit of analysis. To this end Horwich makes a distinction between *strong* determination and *weak* determination. Strong determination is of the type just considered, whereby truth can be analysed in terms of, and hence 'read off', use. The alternative account - weak determination - does not require that truth be 'read off' from use, but merely that "two predicates with the same meaning-constitutional property are co-extensional" (Horwich 1995, p. 363). In particular, with weak determination there is no means of saying which use determines which truth-conditions (in Horwich's terminology, you cannot 'read off' the truth-conditions from the use), only that the same use gives the same truth-conditions.

The distinction Horwich draws between strong and weak determination is legitimate, for weak determination is a respectable notion of determination. For example, in saying that the chemical properties of an element are determined by the number of protons in the nucleus, it is certainly meant that the same number of protons gives the same chemical properties. What does not follow is that there is any means by which the chemical properties of an element can be ascertained simply on the basis of its atomic number.

The fundamental advantage of saying that use weakly determines truth - i.e. same use, same truth-conditions - is that even though we lack the means to correlate specific uses with specific meanings, we still have an informative determination statement. With weak determination there is no requirement that the identity of the truth-condition be ascertained from the use, and so no need to give an analysis of truth in terms of use.

So far, everything is in order: weak determination does appear to give a sense of truth-determination which is consistent with the use theory. To clinch the argument, though, Horwich must do more than establish consistency; he must show not only that deflationary truth is compatible with the notion of weak determination, but also that use actually does weakly determine truth. For his part, Horwich finds the claim that use weakly determines truth to be "uncontroversial" (1995, p. 363), and thus not worthy of serious justification.[4] However, it transpires that the matter is far from uncontroversial. For if we cannot 'read off' an extension from a given use, how can say that the same use determines the same extension? Indeed, what right can we then have to say that an extension is determined at all?

The problem can be illustrated in terms of our example of atomic numbers. The standard strategy to establish a weak determination relation would be to take various samples, identify the respective atomic numbers, test for a range of chemical properties, and see if there is any correlation. Whilst this method does not allow one to 'read off' the chemical properties from the atomic number of an untested element, the process does requires that you be able to tell (empirically) which atomic number is associated with which chemical properties. So although in this case we cannot give a function from atomic number to chemical nature, if we could not in some sense 'read off' the chemical properties of an element from its atomic number (in this case empirically), then the weak determination claim could not be justified. It is notable that a similar process cannot be applied in the case of use and truth, for by assumption, a truth-condition cannot be ascertained from a specified use. To establish weak determination in this case, a more inventive strategy is required.

Establishing Weak Determination

Horwich does not provide the necessary detail, but it is not difficult to see how the account must go, for we have limited materials at our disposal - namely the use theory of meaning

---

[4] Horwich does defend the claim that use weakly determines truth to some extent, but only in as much as it is threatened by Kripke's observation that dispositions are finite. A finite disposition cannot determine an infinite extension, and so dispositions appear to leave meaning underdetermined - in which case same use would not determine same meaning. (It would seem that, as far as finite use can determine, it is possible that two people share the same dispositions, but one means plus, whilst the other means quus.) However, as Horwich observes (1995, p.366) this argument begs the question in assuming that meaning extends beyond use. If we have an independent reason to reduce meaning to use (which the use theorist must have), then there can be no grounds on which to say that two people who use a word in the same way mean different things. Horwich's aim is to counter the objection that use does not determine truth-conditions to within uniqueness. Yet the claim in most need of support - but which Horwich merely recommends as being 'uncontroversial' - is not whether use (weakly) determines an extension to within uniqueness, but whether use (weakly) determines an extension at all.

and the deflationary theory of truth. However, before constructing the argument it is necessary be more precise in our formulation of the deflationary theory.

In the paper under consideration (Horwich 1995), deflationism is characterised as the claim that the disquotational schema:

'p' is true $\Leftrightarrow$ p

captures the notion of truth. The intended function of the quotation marks on the left hand side is to turn a proposition into the name of a sentence. Yet a sentence only has a truth-condition given that it has the meaning that it does. In the above formulation, the fact that 'p' means that p is a tacit assumption. In the normal course of things such an assumption is quite acceptable, but here the aim is to make the nature of truth fully perspicuous, in which case such tacit assumptions have no place. As given, the deflationary schema misses out what is most important: that a sentence has a truth-condition only because the sentence means what it does.[5]

A better strategy is to formulate deflationism in either in terms of propositions:

(ES) <p> is true $\Leftrightarrow$ p

('<p>' stands for 'the proposition that p'); or in terms of utterances:

(DS+) u says that p $\Rightarrow$ (u is true $\Leftrightarrow$ p)

Whilst Horwich gives his sustained defence of deflationism in terms of ES, he also claims (1990, pp. 106-107) that the deflationary theory can equivalently be given in terms of

---

[5] To put the point more forcefully it could be argued that DS gives a vacuous formulation. I am prepared to endorse both of the following conditionals:

(In English) "Bill wears suspenders" is true iff Bill wears suspenders.

(In American English) "Bill wears suspenders" is true iff Bill wears braces.

Do I then endorse the following instance of DS:

"Bill wears suspenders" is true iff Bill wears suspenders?

Clearly there is no correct answer to this question - if the quoted sentence has its English meaning I do, and if it has its American English meaning I do not. There is then no fact of the matter whether I accept this instance of

utterances. This form (i.e. DS+) is the most convenient when it comes to showing that use weakly determines truth, and so it is the one I shall concentrate on.[6]

Given this formulation of the deflationary schema, the fact that use weakly determines truth is readily established. We start with the use theory of meaning. Suppose that S has a given use **u**, and that **u** determines that S means that p. That is:

(1) $U(S) = \mathbf{u} \Rightarrow$ S means that p

where 'U(S)' stands for the use of the sentence S.[7]

The deflationary theory of truth gives us:

(2) S means that p $\Rightarrow$ (S is true $\Leftrightarrow$ p)

DS, the matter is underdetermined. For this reason, DS cannot capture truth. This type of consideration - that the same sentence may have more than one meaning - is acknowledged by Horwich (1990, p. 104).

[6] Horwich actually gives the deflationary theory for utterances using the following schema:

(D-tr) u is true $\Leftrightarrow$ p

"where 'u' is replaced by a singular term referring to an utterance and 'p' is replaced by a sentence of our language that...would be the...translation of that utterance." (Horwich 1990, p. 106). The formulation of the schema D-tr is, however, disingenuous, for the qualification given in text following D-tr (i.e. "where 'u' is replaced by a singular term..." (quoted above)) is really part of the theory itself. The situation is more perspicuous if we include the qualification as an antecedent, giving the following schema:

(D-tr+) trans(u) $\in$ 'p' $\Rightarrow$ (u is true $\Leftrightarrow$ p)

But this formulation (also) rests on the false assumption that 'p' has its content essentially ('p' may mean anything in some possible language). It is for this reason that the formulation in terms of utterances given above, namely DS+, is preferable.

Horwich ought to find the characterisation of deflationism in terms of DS+ wholly unobjectionable. He claims (1990, pp. 106-108) that his own versions of deflationism in terms of utterances (i.e. D-tr+) and propositions (i.e. ES) are *interderivable* given the following 'auxiliary assumptions':

u expresses <p> $\Leftrightarrow$ trans(u) $\in$ 'p'

u expresses <p> $\Rightarrow$ (u is true $\Leftrightarrow$ <p> is true)

If the deflationary theory is given as suggested above, namely:

(DS+) u says that p $\Rightarrow$ (u is true $\Leftrightarrow$ p is true)

then given the auxiliary assumption:

u expresses <p> $\Rightarrow$ (u is true $\Leftrightarrow$ <p> is true)

which Horwich *already* endorses, the deflationary theory given in terms of ES can be derived from the theory given in terms of DS+, and *vice versa*. Since DS+ is equivalent to ES given the accepted auxiliary assumption, Horwich should accept the DS+ formulation.

[7] An alternative approach would be to consider the uses of individual words. The form of the argument would of course be similar.

By transitivity on (1) and (2) we get:

(3) $U(S) = \mathbf{u} \Rightarrow (S \text{ is true} \Leftrightarrow p)$

Of course nothing in the argument is particular to S. By substitution T for S in (1) and (2) the same reasoning gives us:

(4) $U(T) = \mathbf{u} \Rightarrow (T \text{ is true} \Leftrightarrow p)$

A logical consequence of (3) and (4) is:[8]

(5) $[(U(S) = \mathbf{u} \;\&\; U(T) = \mathbf{u})] \Rightarrow [(S \text{ is true} \Leftrightarrow p) \;\&\; (T \text{ is true} \Leftrightarrow p)]$

which in turn entails:

(6) $(U(S) = \mathbf{u} \;\&\; U(T) = \mathbf{u}) \Rightarrow (S \text{ is true} \Leftrightarrow T \text{ is true})$

The choice of **u** is clearly arbitrary, so on condition that S and T share the same meaning-determining use, they will have the same truth-conditions. The conclusion is, then, that use does indeed weakly determine truth.

Against Deflationism

The question I want to raise with respect to this argument is this: what right do we have to use the premise

S means that $p \Rightarrow (S \text{ is true} \Leftrightarrow p)$?

First note that this step is essential, for we start with a statement about meaning, and aim to establish a conclusion about truth-conditions, so at some stage we have to connect meaning and truth. The only possible means of doing this is with an instance of the deflationary schema DS+.

---

[8] Explicitly: $A \Rightarrow B,\ C \Rightarrow D \mid (A \;\&\; C) \Rightarrow (B \;\&\; D)$.

In normal circumstances, DS+ is taken as a platitude, an indisputable fact about truth, and something which is a common starting point for any theory of truth. In the present context, though, this conditional cannot be taken for granted, for what it states is that meaning determines truth, which is precisely the point in question. Rather than simply assuming that meaning determines truth, the deflationist has to *justify* this claim.

The justification of any instance of DS+ may appear to be an entirely trivial matter, for the deflationary thesis, accepted by hypothesis, consists in the claim that DS+ (or equivalently ES) is meaning-determining. As Horwich states:

> The disposition to assert, *a priori*, all instances of '<p> is true iff p' constitutes an *implicit (meaning giving) definition* of the truth predicate. (Emphasis added. Horwich 1990, p. 23, fn)

The deflationist claim about the meaning of 'true' is made against the background of a use theory of meaning. Any claim about the meaning of 'true' has to be reconciled with the adopted position that the meaning of a word is determined by the use to which it is put. It is no accident, therefore, that Horwich gives the following account which combines deflationism with the use theory:

> A person's understanding of the truth-predicate, 'is true' - his knowledge of its meaning - *consists in his disposition to accept*, without evidence, any instantiation of the schema
>
> (E) 'The proposition *that p* is true if and only if p'
>
> by a declarative sentence of English. (Emphasis added. Horwich 1990, p. 36)

In the quoted passage, Horwich refers to his preferred formulation of deflationism in terms of propositions, but given that deflationism can also be formulated in terms of sentences and utterances, the same thought ought to hold in terms of DS+. Correspondingly, DS+ may also be considered a *definition* of the truth predicate. What better justification for DS+ could there be?

On inspection, though, this argument falls apart. The aim is to establish that any instance of DS+ has the status of an established premise, thus making it available in the above demonstration that use weakly determines truth. The basis on which this is done is (a) the fact that meaning is use, and (b) the fact that a significant use of 'true' is our disposition to assert instances of DS+ (or ES). Combining these latter two points, it appears reasonable to hold that the disposition to assert instances of DS+ does indeed give 'true' its meaning.

Yet, expressing the idea in a slightly different way uncovers its difficulty. Let '<u>DS+</u>' be a sentence expressing a particular instance of the schema DS+, and let <u>DS+</u> be the proposition expressed by that sentence. In that case, the claim that our disposition to assert instances of DS+ is meaning-determining, and that this makes any instance of DS+ available as a premise for our argument, is the claim that the following conditional is true:[9]

(*) U('<u>DS+</u>') ⇒ <u>DS+</u>

That is to say, our disposition to assert any instance of DS+ without justification entails that the proposition that that instance expresses is true.

This would be a remarkable claim to make, for the antecedent concerns a linguistic fact, namely that we utter a certain sentence in certain situations, whereas the consequent records a logical relation between meaning and truth, something which should hold irrespective of the contingent use of any sentence. This cannot be right, so something must be amiss.

A precise diagnosis of the error made in (tacitly) endorsing the conditional (*) can be made if we go back a few steps. Under the use theory, one thing that does follow from the use to which '<u>DS+</u>' is put it that it has the meaning that it actually has. That is:

(1) U('<u>DS+</u>') ⇒ '<u>DS+</u>' means that <u>DS+</u>

In saying that our disposition to assert '<u>DS+</u>' is meaning determining, Horwich might mean that this use makes the sentence true (after all, a definition is usually considered to determine a truth by stipulation). If so, the claim is:

(2) U('<u>DS+</u>') ⇒ '<u>DS+</u>' is true.

Employing the trivial premise:

---

[9] Spelling out a specific example of the conditional gives the cumbersome:

U("'Grass is green' means that grass is green ⇒ ('Grass is green' is true ⇔ grass is green)") ⇒

('Grass is green' means that grass is green ⇒ ('Grass is green' is true ⇔ grass is green)).

(3) S means that p, and S is true $\Rightarrow$ p

and with appropriate substitutions, (1) and (2) entail:

(4) U('DS+') $\Rightarrow$ DS+

which is the desired result.

The problem with this argument is not hard to find. Obviously step (2) requires that the use of a sentence can determine that that sentence is true. But in taking this step, we help ourselves to truth by stipulation, which is to say that we *assume* that use determines truth. Since this is precisely the issue under investigation - *can* use determine truth? - the argument clearly begs the question.

Naturally, the above is only one attempt to construct the type of argument that Horwich's thesis must rely on. Nevertheless, there is no need to consider other approaches Horwich might take to reach the desired goal, for we are now able to see that the very basis of Horwich's position is unsound. The only way that the deflationary theory can make the desired connection between use (i.e. meaning) and truth is by appeal to particular instances of the deflationary schema (DS+), the idea is being that instances of the deflationary schema hold trivially, by definition, and are thus available to everyone, no matter what theory of meaning they endorse. But, of course, instances of the deflationary schema can only be used to deliver trivial truth-conditions if they themselves are *true*. This gives rise to the fundamental tension in the position, for deflationism is at root a claim about the *meaning* of the word 'true', and so the deflationary theory itself comes under the jurisdiction of the use theory. So unless our use of 'true' can determine that instances of DS+ are true - unless there is truth by stipulation - then the deflationary theory is itself powerless to establish that any instance of the deflationary schema is true. In short, deflationism only delivers truth-determination if use can be truth-determining, which makes the argument inherently circular.

An Alternative Strategy

There is an alternative approach which anyone wishing to defend Horwich's position may take. Recall that Horwich claims that the disposition to assert instances of the deflationary schema without evidence is constitutive of grasp of the concept *true*. It follows, of course, that anyone who does not have this disposition does not have the concept. And in that case

there can be no significant dispute over whether use determines truth: anyone who does not accept every instance of DS+ is someone who does not have the concept, and so anyone persuaded (perhaps by the indexical argument) that meaning does not determine truth is not wrong, they have simply changed the subject. In that case no one who has the concept true will deny any instance of DS+, and so will have to accept the argument given above that meaning does weakly determine truth.

As it happens, this argument is also defective, for the boot is on the other foot. At root, the central claim is that for some p (i.e. DS+), if you are not disposed to say that p come what may - even in the face of a prima facie argument that not p - then you do not have the concepts necessary to express p. But this is just an act of stonewalling, for this method could be adopted by the use theorist in any situation, to defeat any revisionary argument.

To bring the point out, note that the aim of the argument is to persuade us that, contrary to first appearance, use is a normative, truth-determining, property. That is, deflationism is employed to *change our opinion* about meaning, use and truth. But suppose that my initial reaction is to accept my (reasonable) first impression, namely that since use does not determine truth, and given that meaning is use, then meaning does not determine truth. The job the deflationist theory has is to persuade me that my first impression is in error, and he tries to do this by claiming that anyone who knows what 'true' means will not accept the *prima facie* argument, and will instead assert instances of DS+ *come what may*.

But by hypothesis, I would reject an instance of DS+ were it not for the deflationist's argument. And the only way deflationism can make a difference is if previously I did not have the disposition to assert instances of DS+ come what may. According to the deflationist, this means that I did not have the concept *true* - or at least not the concept of truth captured by the deflationary theory. But if I *change* my verdict in light of the deflationist's argument - if he succeeds in altering my opinion - then by the same token I have changed my concept. So whereas I previously did not have the (deflationary) concept *true* (in virtue of the fact that I was not prepared to assert an instance of DS+), I now do.

Yet an argument ought not convince anyone to change their concepts, but only alter the way they apply their existing concepts. So I will only be persuaded to alter my opinion if I think that the concept of truth I do have is the deflationary concept. In this way deflationism becomes a self-fulfilling theory: because I think that deflationism is true, I alter my verdicts

in such a way that makes it true. Because I think that those who have the concept of truth are disposed to assert DS+ come what may, I assert DS+ come what may. But this process is a complete repudiation of rationality: we ought not alter our concepts to sustain our theories about them, our theories ought to describe what our actual concepts are. In short, the use theory ought to respect our practice of altering what we say in the light of new information, and correspondingly ought not stand in the way of a rationally supported overhaul of our thoughts about truth. In light of this obvious failing, this defence of deflationary truth-determination cannot succeed either.[10]

Other Deflationary Theories

The question we have been concerned with so far is whether Horwich's deflationary theory in particular can make use truth-determining, and our answer is that it cannot. Nevertheless, Horwich's theory is only one version of deflationism, and so the additional question arises whether some other deflationary thesis might prove more successful. One notable point is that Horwich's version of deflationism maintains that truth is a property, whereas a more traditional characterisation - indeed a definition - of deflationsim is simply the denial that truth is a property at all.[11] The type of theory in mind here is a type of expressivism, so that the sentence "S is true" is not descriptive, but rather expresses a certain type of approval towards S. Could not this type of theory make the ascription of truth to sentences legitimate?

Happily, it is unnecessary to investigate the (somewhat dubious) merits of such a theory in any detail, for the communal use theory cannot support any such thesis. This is because this type of deflationary thesis is normative (as discussed in Chapter 4), in that it holds that "is true" *ought* to be uttered whenever the speaker has the appropriate attitude, and withheld otherwise. In short, the theory is a claim about the rules which govern the use of the predicate "is true". Since rule-following is impossible, the expressive thesis is immediately untenable.

---

[10] This response to Horwich highlights what I believe is a substantial advantage for the communal use theory over a standard dispositional theory. According to Horwich, there are certain identifiable uses of a word - such as the assertion of instances of DS+ - which are meaning-determining. In that case, such uses are enshrined: contravene the meaning-determining use, and you lose the concept. In contrast, the communal use theory does not enshrine any particular uses, and under it the meaning of a sentence cannot be given in terms of specific unassailable uses. All we can say is that, broadly speaking, those who use a word in the same way mean the same thing. Hence such agreement might involve the rejection of firmly held beliefs (even those perhaps previously asserted without the need for justification), should the appropriate refutations arise.

[11] See for example Kirkham (1992).

Eliminating Truth

Communal use can neither strongly nor weakly determine truth; under the use theory "is true" cannot be expressive. There is no other option but to accept that communal use does not determine truth at all. Although strongly counter-intuitive, there is no option but to endorse truth-elimination. In that case, the immediate task is to show that the removal of truth does not lead to an unsustainable world view.[12]

To see whether the loss of truth can be sustained we really need to know what we lose when we lose truth, or, put the other way round, to identify why we think we need truth in the first place. In asking what we need truth for, there are really two issues to consider:

(a) Why do we need the concept *true*?
(b) Why do we need anything to have truth-conditions, or to be true?

We need the concept *true*, fairly obviously, so that we can say things which otherwise could not be said (or to think things which otherwise could not be thought). Established candidates are generalisations such as "Every sentence of the form 'P or not P' is true", and sentences which allow for the affirmation of a statement without repeating it - 'blind endorsements' such as "What the Bishop said is true".

Yet, no matter which particular statements feature an ineliminable truth-predicate, this issue has no bearing on the question of whether anything actually has truth-conditions or a truth-value. For to enable us to say things which could not otherwise be said in this way, all that is required is that 'is true' has a distinct meaning, which on the use theory is to demand only that it has a distinct use.[13]

---

[12] The complete elimination of truth would require the claim not only that no linguistic item is true, but also that no mental state such as a belief, nor abstract object such as a proposition, can be true either. Above I concentrate on the failure of truth to adhere to sentences and utterances, but the expectation has to be that mental content receives a similar treatment. As for propositions, to say that a sentence or utterance is not true means that what it expresses is not true. We are then faced with two options: either sentences and utterances express propositions, and propositions are not truth-bearers; or sentences and utterances do not express propositions, in which case propositions could be truth-bearers. In the latter case, propositions should cease to hold any interest for us - indeed, since propositions are posited precisely in order to account for the common content of disparate sentences, we should then have no reason to suppose that such things exist. So either way, we can eliminate propositions as truth-bearers. Clearly a satisfactory treatment of these issues would require far more discussion than there is room for here, but the prospect is that the thesis as formulated with respect to linguistic items does indeed entail the complete elimination of truth.

[13] It is sometimes claimed that truth is required solely to allow for such generalisations (e.g. Quine 1970, pp.11-12) and blind endorsements. It would be fine for the 'no truth' theorist if this were so, for the fact that the concept of truth is required in order to make a certain type of statement does not entail that anything has the property of

The more important question in the present context is, therefore, the second one: why do we need anything to actually have a truth-condition, or a truth-value? Apart from (supposedly) accounting for content, a sample of the phenomena which plausibly require truth to be adequately explained might include:

(a) the nature of assertion
(b) the success of rational (scientific) enquiry
(c) the validity of certain forms of reasoning.[14]

Importantly, it is not just our view of ourselves as agents who engage in such activities which is at stake here. For the 'no truth' theory is, naturally, supposedly asserted on the basis of rational enquiry using valid reasoning. If any one of these activities turns out to be unavailable precisely because nothing is true, then clearly the position would be self-defeating.

One point worth making at the outset is that not every explanation in the above list requires that any linguistic item actually have truth-conditions or a truth-value. To give an example, in saying that truth is required to explain the nature of assertion, what is often meant is that truth is a norm on the practice of assertion. That is, to assert is to aim at the truth. But to aim at the truth - to assert what you believe to be true because you believe it to be true - does not depend on anything actually being true. All that is needed is that you believe what you assert to be true. So again, it is only the concept of truth, and not the instantiation of truth, which is essential to the explanation, and this is quite acceptable to the 'no truth' theorist.[15]

---

truth. (Similarly, the concept of *witch* is necessary to describe someone as a witch, but that hardly entails that anyone actually is a witch.) Hence it is quite consistent to say that truth is (only) necessary to make blind ascriptions and certain generalisation, as well as claiming (as the 'no truth' theorist does) that no sentence of the form "P or not P" is true, and that in all cases what the Bishop said is not true.

[14] Instances where each type of claim is made are: (a) Dummett (1976, p. 83) and Williams (1973, p. 202); (b) Putnam (1978); (c) Dummett (1977) and Putnam (1978).

[15] How can someone who accepts the 'no truth' theory, and who thereby believes that truth is unobtainable, still aim at truth when making an assertion? Obviously there is immense practical benefit to be had from engaging in the practice of assertion, and so it is perfectly rational to suspend one's belief, and act *as if* one believed in truth. We should no more say that such a person is not actually making an assertion than we should say that someone who knows he cannot win a game, but continues to act as if he can, is not really playing the game; or that someone who tells a lie does not thereby make an assertion. In other words, to engage in the practice of assertion is not necessary to aim at truth, but merely to act as if aiming at truth.

I shall not attempt to decide whether this strategy can be applied universally, for the more important point is that even if the possession of truth-conditions *is* essential for some explanatory task (whether mentioned on the list or not), this type of concern cannot be used to overturn the 'no truth' theory. The position we are considering is one in which the following propositions are taken as established:

(1) Nothing can have truth-conditions.
(2) The explanation of X requires that something have truth-conditions.

The threat to the 'no truth' theory arises initially because it is now deemed to be explanatorily inadequate, and (more forcefully) on the basis that X (a certain inferential process, say) may be required to establish the theory in the first place.

But to argue in this manner is to overlook the fact that there is no guarantee that everything we would like an explanation for can actually be explained. Basically, we are faced with a choice: given that X cannot be explained in the terms available, we either accept that X is impossible, *or* we accept that X is inexplicable. If X is some process which really is indispensable, on the basis that it is necessary for rational enquiry and so cannot be eliminated on the basis of reason, then the latter option - that X cannot be explained - is the option to take. In short, we can never overturn an argument that nothing can be F on the basis that F is needed in some explanation. Explanations, though clearly desirable, can never be demanded as of right.

This point is really a reiteration of the argument given in Chapter 6. For it may be that normally, in as much as we have reason to believe that people engage in activity X, we also have reason to believe that something has a truth-condition. Consequently, we have reason to believe that X can only be performed if possession of a truth-condition is possible. Yet, once it is shown that grasp of a rule, and hence possession of a truth-condition, is impossible, this ought to motivate the claim, not that X is impossible, but only that X is not after all to be explained in the manner thought. Whilst in the case of content, a replacement notion was identified to take the place of rules/truth, there is nothing to ensure that a replacement notion is always available. And even if there is not, the outcome is not altered. In such a case, we should still conclude that X is not to be explained as previously thought, but given that we have as much reason to believe that people engage in the activity as we ever did, the only option is to conclude that X admits of no substantive explanation. Explanatory deficiency is

not a logical difficulty, and so there can, then, be no argument that the 'no truth' thesis is incoherent along these lines.

A Final Rejoinder

There remains, however, the prospect of a genuine incoherence. The 'no truth' theory appears to be inherently unstable: if no sentence is true, then the sentence "No sentence is true" is itself not true. Does this make the thesis self-defeating?

It does not, and the explanation is brief. The argument, in slightly more detail, is that given the 'no truth' thesis, namely:

(1) $\forall$S, S is not true

we can substitute the sentence in (1) as an instance of S in (1) itself to get:

(2) "$\forall$S, S is not true" is not true.

Given that:

(3) 'p' is true $\Leftrightarrow$ p

then by MTT on the right-to-left conditional of (3) we have:

(4) Not: $\forall$S, S is not true

which contradicts (1).

The problem here is by now familiar, for in using the 'disquotational' property of truth, it should not be forgotten that the meaning of the mentioned sentence is left tacit. That is:

"$\forall$S, S is not true" is not true $\Rightarrow$ Not: $\forall$S, S is not true

relies on the contingent fact that the sentence "$\forall$S, S is not true" *means* that no sentence is true. So the 'disquotational' principle actually involved is:

S means that p $\Rightarrow$ (S is true $\Leftrightarrow$ p),

that is, DS+. Previously it was argued that the use theorist has no right to this platitude. (In particular, the minimalist claim that such a platitude gives the meaning of 'true' was rejected, for under the use theory it is never possible to encapsulate the meaning of a word in this way. All we can ever say is that knowledge of meaning consists in possession of a certain kind of ability.) The 'no truth' theory was adopted precisely because, without DS+ as a premise, it is not possible to get truth from meaning. That is, it is because the claim that meaning determines truth cannot be justified that the truth-eliminativist accepts that meaning does not determine truth. Since DS+ just is the claim that meaning determines truth, the step of adopting truth-elimination consists in the rejection of DS+ on the grounds that it is unjustifiable.[16] Hence the truth-eliminativist has *already* rejected DS+, and so it cannot be used as a premise in an argument against his position.[17]

---

[16] Since DS+ is a statement which most people would accept, the 'no truth' theorist's rejection of it is certainly revisionary of our ordinary linguistic practice. Such revisionism is not a problem for the use theorist here described. For, as discussed above, the theory does not hold that the endorsement of any particular utterances (even though they may enjoy universal communal assent) is part of what it is to know a particular meaning. Rather, one must be able to respond to various stimuli, in an on-going fashion, in broadly the same way as others would when in the same situation. Hence, if your stimulus is different from that of others, it is to be expected that the things you say will also be different; and if you absorb new arguments, it is to be expected that you will alter you theoretical position accordingly, even if this puts you out of step with what others currently say. Hence, the rejection of various truth-platitudes (as enjoined by truth-elimination) as the result of rational reflection is consistent with the engagement in the communal use of the word 'true'.

[17] The 'no truth' theory of truth is not without historical precedence, having been advocated by Nietzsche (1873/1979, p. 84).

Conclusion

As things stand, the communal use theory is far from fully developed. Indeed there are many pressing questions which would have to be answered for the theory to be anything like complete. In particular, though I cannot address any of them here, there are several issues arising in relation to the nature and type of the agreement, and the classification of communal practices, required by the theory. To give a sample:

- In its present form the theory describes content in terms of agreement in judgement (as to what is correct). Is a theory of content which makes essential reference to mental content in this way really a stable position? (Specific charges that any reference to content within a theory of content leads to vicious circularity were considered, and dismissed, in Chapters 2 and 5, but the position remains somewhat uncomfortable. Can it be made more secure?)

- To agree is to give the same verdict in the same context, so that the existence of agreement would appear to depend upon our existing world view, that is on the way that we classify different contexts. Does this make agreement, and hence meaning, somehow subjective?

- In talking about agreement there is a *prima facie* need for publicity - we not only have to agree, but have to be *seen* to agree. How does such a publicity requirement have bearing on the occurrence of agreement? Does the need for verdicts to be manifestable in behaviour give rise to fresh holistic concerns?

- The communal use theory states that meaningful employment of words requires, not that we actually follow rules, only that we *look* like we are following rules. Is it then possible to say *which* rules you have to look like we are following - which practices we have to agree in - for our words to have meaning? (Clearly it cannot require straightforward agreement in application, for there are many subject matters - such as aesthetics and morals - where such agreement is rare.) Can we say specifically which practices we have

to engage in to count as knowing the meaning of a specific word, and if so would such result hold any significance?

Despite these unresolved issues, it is plausible that we already have sufficient detail to discern (if somewhat provisionally) the general philosophical message of the communal use theory, and it is in outlining this that I should like to end. Unfortunately, it turns out that the only message it is possible to discern at the present time is that the philosophical message of the communal use theory is hard to discern.

In ascertaining the metaphysical significance of any semantic theory, we need first to consider what metaphysics is. Traditionally, the central aim of metaphysics is to explain (as far as possible) the sensible world in terms of the suprasensible. To this end we should (a) identify the various entities in the world required for this explanatory task and (b) uncover the various relations holding between them. As a result, we can expect a metaphysical theory to include (a) a set of existence claims, and (b) a set of conditionals recording the logical relations existing between the items in that ontology. To take a typical example, we might explain the behaviour of others in terms of various mental states (and so posit the existence of other minds, beliefs, pains, and so on), and hope to capture the nature of such states by recording the logical relations holding between them (noting that if x is a belief, it is not a desire; that if S is in pain, she knows it; and so on.)

The connection between these metaphysical aims and a semantic theory arises from the dependency which exists between the ability of the relevant statements to perform the duties required of them by a metaphysical theory, and the way in which such statements are warranted. In particular, the two different types of statement we are concerned with (existence claims, and conditionals) must each perform quite specific tasks. As noted above, we expect existence statements that are part of a warranted metaphysical thesis to carry a certain *explanatory* burden; and we expect those conditionals which form part of a warranted metaphysical thesis to have the capacity to convey *information* about the way of the world. Whether either function - explanatory power, and informativeness respectively -can be sustained will depend on the mechanism which bestows the relevant warrants on our assertoric and inferential practices, and that is a question for our theory of meaning.

To put this point in context, let us consider theories of meaning based on the following two notions:

(a) Truth-conditions;

(b) Conceptual role.

The truth-conditional theory states that knowledge of meaning consists in grasp of truth-conditions, and that on the basis of this knowledge we are able to *derive* knowledge of what warrants the assertion of a particular statement S, and also how S may feature, either as premise or conclusion, within various inferences (i.e. its conceptual role). This theory has a high degree of explanatory unity, for both assertoric and inferential practices are derived from one key concept.

In contrast, a conceptual role theory (or at least one version of a conceptual role theory) relinquishes the idea that both assertoric and inferential practices may derived from a common source: instead, assertion- and inference-conditions are themselves basic constituents of meaning. That is, grasp of meaning consists in grasping a set of appropriate rules governing assertoric and inferential practices.

In light of the different ways that they account for our linguistic practices, each theory entails a different metaphysical view. Taking the truth-conditional theory first, under this theory in order that a statement such as "S is in pain" be warranted, we must have reason to believe that its truth-condition is satisfied. Clearly, the fact that pain behaviour warrants the relevant claim, and so indicates that the truth-condition for "S is in pain" is satisfied is not something which is determined purely by the meaning of that expression. Rather, it is a matter determined by the way of the world. Specifically, it must be that the fact that the truth-condition is satisfied - that is to say, that S is in pain - is the best explanation for the existence of her pain behaviour. In this way, the warrant in question must rest on inference to the best explanation, and inference to the best explanation is only possible if the existence claim in question is genuinely explanatory.

In addition, if meaning is truth-conditional, then since we know what makes "S is in pain" true, and what makes "S has a mind" true, we are in an excellent position to recognise that the truth of the former guarantees the truth of the latter. And since the conditional "If S is in pain, S has a mind" records a relation between truth-conditions, it thereby records a relation which must hold between those entities which satisfy those truth-conditions. In this way conditionals do indeed serve to record relations between those properties instantiated in the

world, and can indeed be informative, should the relevant properties actually be instantiated. As a result, the truth-conditional theory accounts for our assertoric and inferential practices in a way which allows them to be (respectively) explanatory and informative, which is to say this theory of meaning is compatible with a substantive metaphysics.

The same cannot be said about the conceptual role theory. If assertion-conditions are constitutive of meaning, then the fact that pain is ascribable on the basis of behaviour is not made on the basis of inference to the best explanation. Rather, the assertion is warranted directly on the basis of an arbitrary semantic norm. We cannot expect an arbitrary norm to increase (arbitrarily) our ability to explain the world, in which case talk of pain cannot be used to explain the existence of pain behaviour.

Likewise, if knowledge of conceptual role is an independent constituent of meaning, then statements of the type of *a priori* conditionals philosophers are concerned with do not record relations between the relevant truth-makers, but merely record the arbitrary, meaning-giving rules governing the use of the terms involved. Consequently, conditionals recording warranted inferences in no way reflect the underlying nature of reality.

Under the conceptual role theory - a view plausibly ascribed to Wittgenstein - neither type of element of a 'metaphysical' theory can sustain the role required for it to be genuinely metaphysical. Rather, all that we can hope to do is to is to identify the set of rules (the 'grammar') governing the various assertoric and inferential practices which govern a discourse.[1] In that case it is the fact that different subject matters are governed by different use-rules - they constitute different language games - that leads us to think that there are distinct types of entity in the world with differing ontological status. It is the mistake of the metaphysician to try to account for what are really mere differences in 'grammar' in the more substantive terms of a metaphysical system, the error being that of seeing the differences in question as having significance beyond the semantic. Once this error is recognised, metaphysics collapses.

The reason for examining these two theories is that they provide a framework within which it should be possible to place the communal use theory. However, this placement is not straightforward, for there are two competing forces at work which are not readily reconciled.

---

[1] Cf. Zettel §590

To start with, we could consider the communal use theory as the result of moving in a direction taking us away from a truth-conditional theory and towards a conceptual role theory, and then taking an extra step beyond. This is because the three theories - truth-conditions, conceptual role, communal use - are readily put on a scale of decreasing explanatory power. As we have seen, the truth-conditional theory explains both the principles governing assertion and inference in terms of one key concept, namely truth. The conceptual role theory is reached by removing that unifying explanatory layer, leaving only the 'surface' rules governing use to fulfil all explanatory roles. In turn, the communal use theory states that we cannot even appeal to these types of rule to explain our linguistic practices. Rather than consisting in the grasp of a set of rules governing the use of our words, grasp of meaning consists in the ability to engage in various assertoric/inferential practices, where such abilities are taken as basic, and not amenable to explanation at all. Initially, then, linguistic practices are claimed to be governed by a single type of underlying rule, then they are held to be governed directly by independent use rules, and finally by no rules at all. In this way the progression from truth-conditions to conceptual role to communal use is the result of removing one layer rules - and hence one layer of explanatory depth - at each stage.

When approached from this direction, it appears that at the (limited) level of explanatory power provided by a conceptual role semantics we have already relinquished the possibility of metaphysics, and so making the further move to communal use merely serves to make things (from the point of view of the metaphysician) worse. In particular, under the communal use theory our relevant assertoric and inferential practices are not warranted by anything - the practices are just what we do - and so certainly cannot accommodate the type of warrant required for an explanatory metaphysics.[2] So, in making the step from conceptual role to communal use, the lot of the metaphysician is in no way improved, and from this perspective it looks as though communal use is as destructive of a substantive metaphysics as the conceptual role theory is.

However, before endorsing this conclusion, there is an alternative approach to this issue which should be considered. The above line of reasoning could be summed up as the thought that because it is a truth-conditional semantics which supports a traditional metaphysical

---

[2] Of course when using language we have to consider that our linguistic utterances are warranted. The point is that from the meta-linguistic viewpoint - from outside the language in question - there are no rules available to take on the necessary normative role.

outlook, *and* because both conceptual role and communal use theories are incompatible with a truth-conditional semantics, that neither is suitable for the formulation of metaphysical theories. Yet, as we have seen, the conceptual role account is not destructive of metaphysics *in virtue* of the fact that it relinquishes truth-conditions, but rather because of the nature of the norms used instead. That is, it is because under this theory use is governed by arbitrary rules of assertion and inference that neither of those practices be used to betray the nature of reality. The relevant point is that the communal use theory *also* relinquishes rules governing assertion and inference, and in doing so distances itself from the type of anti-metaphysical view that the conceptual role theory enjoined.

The overall result is that the communal use falls between the two positions discussed: it is neither metaphysically potent like the truth-conditional theory, nor metaphysically impotent like the conceptual role theory. Deciding what we can say in this situation - for there is no obvious third position to adopt - is perhaps the most pressing issue facing anyone advocating the communal use theory, but it is one which presents no obvious means of resolution.

Bibliography

Allen, B. (1989) 'Gruesome Arithmetic: Kripke's Sceptic Replies', *Dialogue* 28: 257-264.

Ayer, A. J. (1954) 'Can there be a Private Language', *Proceedings of the Aristotelian Society* Suppl. Vol. 28: 63-94.

Appiah, A. (1986) *For Truth in Semantics*, Oxford: Blackwell.

Baker, G. P. and Hacker, P. M. S. (1980) *An Analytical Commentary on the Philosophical Investigations Vol. 1 Wittgenstein: Understanding and Meaning*, Oxford: Blackwell.

Baker, G. P. and Hacker, P. M. S. (1982) *Wittgenstein: Meaning and Understanding*, Oxford: Blackwell.

Baker, G. P. and Hacker, P. M. S. (1984a) 'On Misunderstanding Wittgenstein: Kripke's Private Language Argument', *Synthese* 58: 407-450.

Baker, G. P. and Hacker, P. M. S. (1984b) *Scepticism, Rules and Language*. Oxford: Blackwell.

Baker, G. P. and Hacker, P.M.S. (1985) *An Analytical Commentary on the Philosophical Investigations Vol. 2 Wittgenstein : Rules, Grammar and Necessity*, Oxford: Blackwell,

Barwise, J. and Perry, J. (1983) *Situations and Attitudes*, Cambridge, MA.: Bradford Books/MIT Press.

Blackburn, S. (1984a) *Spreading the Word*, Oxford: Clarendon.

Blackburn, S. (1984b) 'The Individual Strikes Back', *Synthese*, 58: 281-301.

Boghossian, P. A. (1989a) Review of *Wittgenstein On Meaning* by C. McGinn, *Philosophical Review* 98: 83-92.

Boghossian, P. A. (1989b) 'The Rule-Following Considerations', *Mind* 98: 507-549.

Boghossian, P. A. (1990) 'The Status of Content', *Philosophical Review* 99: 157-184.

Budd, M. (1984) 'Wittgenstein On Meaning, Interpretation and Rules', *Synthese* 58: 303-323.

Chisholm, R. M. (1982) *The Foundations of Knowledge.* Sussex: Harvester Press.

Coates, P. (1986) 'Kripke's Sceptical Paradox: Normativeness And Meaning', *Mind* 95: 77-80.

Craig, E. (1982) 'Meaning, Use and Privacy', *Mind* 91: 541-564.

Craig, E. (1986) 'Privacy and Rule-Following' in Butterfield, J. (ed.) *Langauge Mind and Logic*, Cambridge: Cambridge University Press, pp. 169-185.

Davidson, D. (1980) 'Actions, Reasons and Causes', in *Essays on Actions and Events* Oxford: Oxford University Press, pp. 3-19.

Dennett, D. and Haugeland, J. C. 'Intentionality' in Gregory (1987), pp. 383-386.

Devitt, M. (1990) 'Transcendentalism About Content', *Pacific Philosophical Quarterly* 71: 247-263.

Devitt, M. and Rey, G. (1991) 'Transcending Transcendentalism: A Response to Boghossian', *Pacific Philosophical Quarterly* 72: 87-100

Dillard, P. S. (1996) 'Radical Anti-Deflationism', *Philosophy and Phenomenological Research* 56: 173-181.

Divers, J. and Miller, A. (1994) 'Best Opinion, Intention-Detecting and Analytic Functionalism', *Philosophical Quarterly* 44: 239-245.

Dummett, M. (1959a) 'Wittgenstein's Philosophy of Mathematics', *Philosophical Review* 68: 324-248. Reprinted in Dummett (1978), pp. 166-185.

Dummett, M. (1959b) 'Truth', *Proceedings of the Aristotelian Society* 59: 141-162. Reprinted in Dummett (1978) pp. 1-24.

Dummett, M. (1973b) *Frege: Philosophy of Language*. London: Duckworth.

Dummett, M. (1975) 'The Philosophical Basis of Intuitionistic Logic', in H. E. Rose and J. C. Shepherdson (eds.), *Logic Colloquium '73,* pp. 5-40. Reprinted in Dummett (1978), pp. 290-318.

Dummett, M. (1976) 'What is a Theory of Meaning? (II)', in Evans, G. and McDowell, J. (eds.) *Truth and Meaning*, Oxford: Clarendon, pp. 67-137.

Dummett, M. (1978) *Truth and Other Enigmas*. Oxford: Clarendon.

Edwards, J. (1992) 'Best Opinion and Intentional States', *Philosophical Quarterly* 42: 21-33

Field, H. (1994a) 'Deflationist Views of Meaning and Content', *Mind* 103: 249-285.

Field, H. (1994b) 'Disquotational Truth and Factually Defective Discourse', *Philosophical Review* 103: 405-452.

Fogelin, R. J. (1987) *Wittgenstein* (2nd ed.), The Arguments of the Philosophers, London: Routledge & Kegan Paul.

Forbes, G. (1984) 'Scepticism and Semantic Knowledge', *Proceedings of the Aristotelian Society*, 58: 223-237.

Ginet, C. (1992) 'The Dispositionalist Solution to Wittgenstein's Problem about Understanding a Rule: Answering Kripke's Objections', *Midwest Studies in Philosophy* 17: 53-73

Goodman, N. (1973) *Fact, Fiction, and Forecast*, Indianapolis: Bobbs-Merrill.

Gregory, R. L. (1987) *The Oxford Companion to the Mind*, Oxford: OUP.

Grice, H. P. (1957) 'Meaning', *Philosophical Review* 66: 377-388

Grice, H. P. (1969) 'Utterer's Meaning and Intentions', *Philosophical Review*, 78: 2-27

Goldfarb, W. (1985) 'Kripke On Wittgenstein On Rules', *Journal Of Philosophy* 82: 471-488.

Hacking, I. (1993) 'On Kripke's and Goodman's Uses of Grue' *Philosophy* 68: 269-295.

Haldane, J. and Wright, C. (eds.) (1992) *Reality, Representation and Projection*, New York: Oxford University Press.

Heal, J. (1989) *Fact and Meaning*, Oxford: Blackwell.

Holtzman, S. H. and Leich, C. M. (eds.) (1981) *Wittgenstein: To Follow a Rule*, International Library of Philosophy, London: Routledge & Kegan Paul.

Horwich, P. (1990) *Truth*, Oxford: Blackwell.

Horwich, P. (1995) 'Meaning, Use and Truth: On Whether a Use-Type Theory of Meaning is Precluded by the Requirement that Whatever Constitutes the Meaning of a Predicate Be Capable of Determining the Set of Things of Which the Predicate is True and to Which It Ought to be Applied', *Mind* 104: 355-368.

Jackson, F., Oppy, G. and Smith, M. (1994) 'Minimalism and Truth Aptness', *Mind* 103: 287-302.

Kaplan, D. (1990) 'Thoughts on Demonstratives', in P. Yourgrau (ed.), *Demonstratives*, Oxford: Oxford University Press, pp. 34-49.

Kirkham, R. L. (1992) *Theories of Truth*, Cambridge MA: MIT Press.

Kraut, R. (1993) 'Robust Deflationism', *Philosophical Review* 102: 247-263.

Kripke, S. (1982) *Wittgenstein on Rules and Private Language*, Oxford: Blackwell.

Maddy, P. (1984) 'How the Causal Theorist Follows a Rule', Midwest Studies In Philosophy 9: 457-477.

Malcolm, N. (1986) *Wittgenstein: Nothing is Hidden*, Oxford: Blackwell.

Malcolm, N. (1989) 'Wittgenstein on Language and Rules', *Philosophy* 64: 5-28.

McDowell, J. (1984) 'Wittgenstein on Following a Rule', *Synthese* 58: 325-363.

McDowell, J. (1992) 'Meaning and Intentionality in Wittgenstein's Later Philosophy', *Midwest Studies in Philosophy* XVII: 40-50.

McGinn, C. (1984a) *Wittgenstein on Meaning*, Oxford: Blackwell.

McGinn, C. (1984b) 'Kripke On Wittgenstein's Sceptical Problem', *Ratio* 26: 19-31.

Miller, A. (1989) 'An Objection to Wright's Treatment of Intention', *Analysis* 49: 169-173.

Millikan, R.G. (1990) 'Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox', *Philosophical Review* 99: 323-353.

Nietzsche, F. (1979) 'Truth and Lie in the Extra-Moral Sense', in D. Breazeale (trans.) *Truth and Philosophy: Selections from Neitzcshe's Notebooks of the 1870's*, Atlantic Highlands, NJ: Humanities Press. (Original work published 1873).

Peacocke, C. (1984) Review of *Wittgenstein On Rules And Private Language* by S. Kripke, *Philosophical Review* 93: 263-271.

Peacocke, C. (1992) *A Study of Concepts*, Cambridge, MA: MIT Press.

Pears, D. (1988) *The False Prison: A Study of the Development of Wittgenstein's Philosophy* (Vol. 2). Oxford: Clarendon Press.

Pears, D. (1991) 'Wittgenstein's Account of Rule-Following', *Synthese* 87: 273-283

Perry, J. (1977) 'Frege On Demonstratives', *Philosophical Review* LXXXVI: 474-497.

Perry, J. (1979) 'The Problem of the Essential Indexical', *Nous* 13: 3-21.

Pettit, P. (1990a) 'The Reality Of Rule-Following', *Mind* 99: 1-21 .

Pettit, P. (1990b) 'Affirming the Reality of Rule-Following', *Mind* 99: 433-439.

Price, H. (1989) *Facts and the Function of Truth*, Oxford: Blackwell.

Puhl, K. (ed.) (1991) *Meaning Scepticism*, Berlin: de Gruyter.

Putnam, H. (1978) *Meaning and the Moral Sciences*, London: Routledge and Kegan Paul.

Quine, W. V. (1970) *Philosophy of Logic*, Englwood Cliffs: Prentice-Hall.

Quine, W. V. (1992) *Pursuit of Truth* (Rev. ed.), Cambridge, MA: Harvard University Press.

Sartorelli, J. (1991) 'McGinn on Concept Skepticism and Kripke's Skeptical Argument', *Analysis* 51: 79-84.

Searle, J. (1969) *Speech Acts*, Cambridge: Cambridge University Press.

Searle, J. (1979) *Expression and Meaning*, Cambridge: Cambridge University Press.

Shoemaker, S. (1963) *Self-Knowledge and Self-Identity*, Ithica, NY: Cornell University Press.

Shogenji, T. (1992) 'The Boomerang Defense of Rule-Following', *Southern Journal Of Philosophy* 30: 115-122.

Shogenji, T. (1993) 'Modest Skepticism About Rule-Following', *Australasian Journal Of Philosophy* 71: 486-500.

Stabler, D. P. (1987) 'Kripke on Functionalism and Automata', *Synthese* 70: 1-22.

Sullivan, P. M. (1994) 'Problems for a Construction of Meaning and Intention', *Mind* 103: 147-168.

Summerfield, D. M. (1990) 'On Taking the Rabbit of Rule-Following out of the Hat of Representation: A Response to Pettit's "The Reality of Rule-Following"', *Mind* 99: 425-432.

Werhane, P. H. (1987) 'Some Paradoxes In Kripke Interpretation Of Wittgenstein', *Synthese* 73: 253-273.

Werhane, P. H. (1987) 'The Constitutive Nature Of Rules', *Southern Journal of Philosophy* 25: 239-254.

Whyte, J. T. (1992) Review of *Truth* by P. Horwich, *British Journal for the Philosophy of Science* 43: 279-282.

Wittgenstein, L. (1953) *Philosophical Investigations*, Oxford: Blackwell.

Wittgenstein, L. (1956) *Remarks on the Foundations of Mathematics*, Oxford: Blackwell.

Wittgenstein, L. (1958) *The Blue and the Brown Books*, Oxford: Blackwell.

Wittgenstein, L. (1967) *Zettel*, Oxford: Blackwell.

Wright, C. (1980) *Wittgenstein on the Foundations of Mathematics*, London: Duckworth.

Wright, C. (1984) 'Kripke's Account of the Argument Against Private Language', *Journal of Philosophy* 81: 759-778.

Wright, C. (1986) 'Rule-Following, Meaning and Constructivism', in C. Travis (ed.), *Meaning and Interpretation*, Oxford: Blackwell, pp. 271-297.

Wright, C. (1987) 'On Making Up One's Mind: Wittgenstein On Intention', in P. Weingartner & G. Schurz, (eds.) *Logic, Philosophy of Science and Epistemology: Proceedings of the 11th International Wittgenstein Symposium*, Vienna: Holder-Pickler-Tempsky, pp. 391-404.

Wright, C. (1988) 'Moral Values, Projection and Secondary Qualities', *The Aristotelian Society*, Suppl. Vol. 62: 1-26

Wright, C. (1989a) Critical Notice of *Wittgenstein on Meaning* by C. McGinn, *Mind* 98: 289-305.

Wright, C. (1989c) 'Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics', in George, A. (ed.) *Reflections on Chomsky*, Oxford: Blackwell, pp. 233-264.

Wright, C. (1989d) 'Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention', *Journal of Philosophy* 86: 622-634.

Wright, C. (1991) 'Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention', in K. Puhl, (ed.) *Meaning Scepticism*, Berlin: de Gruyter, pp. 128-147.

Wright, C. (1992) *Truth and Objectivity*, Cambridge, MA: Harvard University Press.